

Integration of 6D Object Localization and Obstacle Detection for Collision Free Robotic Manipulation

Thilo Grundmann

Robert Eidenberger

Raoul D. Zoellner

Siemens Corporate Technology

Information and Communications

Otto-Hahn-Ring 6

Munich, Germany

thilo.grundmann.ext@siemens.com

Zhixing Xue

Steffen Ruehl

J. Marius Zoellner

Ruediger Dillmann

Forschungszentrum Informatik

Haid-und-Neu-Str. 10-14

Karlsruhe, Germany

xue@fzi.de

Jens Kuehnle

Alexander Verl

Fraunhofer IPA

Nobelstrasse 12

Stuttgart, Germany

jens.kuehnle@ipa.fraunhofer.de

Abstract — The major goal of research regarding mobile service robotics is to enable a robot to assist human beings in their everyday life. This implies that the robot will have to deal with everyday life environments. One of the most important steps towards able service robots is to enhance the ability to operate well in unstructured living environments. In this paper we focus on the integration of object recognition, obstacle detection and collision free manipulation to increase the service robots manipulation abilities in the context of highly unstructured environments.

Keywords: *manipulation, 6D object localization, obstacle detection, stereo vision, 3D time-of-flight camera*

I. INTRODUCTION

The major goal of research regarding mobile service robotics is to enable robots to assist human beings in their everyday life. This implies that such a robot will have to deal with everyday life environments. Today's robots are able to work well in highly structured environments such as fabrication lines however they often have difficulties dealing with the high variability of human living areas. Therefore, one of the most important steps towards able service robots is to enhance the ability to operate well in unstructured living environments.

In the field of mobile service robotics the task of manipulating objects plays an important role. For an able service robot to achieve this central ability a number of subtasks have to be solved. Since grasping can only be planned with respect to a known model, usually a model of the object to be grasped has to be classified and localized. For the planning of a collision free manipulation the robot needs a complete model of the whole volume the manipulation takes place in, independent of whether objects are previously known to the robot or not.

Common systems focus on the recognition and localization of known objects and use these recognition results as basis for the manipulation. Problems arise since in a realistic scenario not only unknown objects are present in the scene, but it may also happen that due to tough conditions known objects are not recognized correctly at all. Since the incorporation of learning

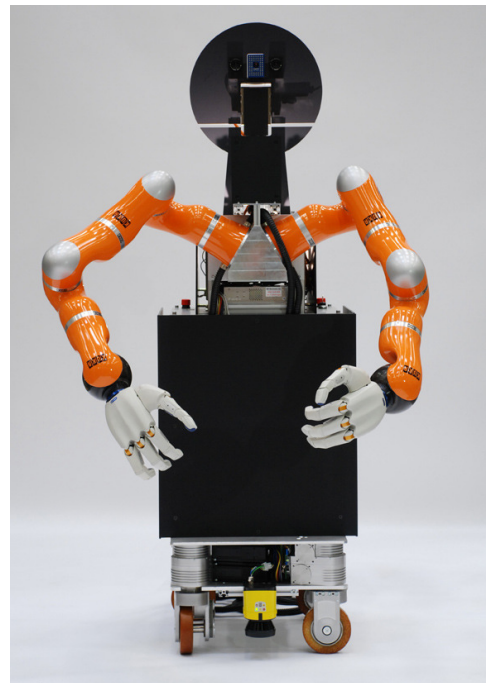


Fig. 1. The DESIRE experimental platform

techniques tackles only the first issue it is necessary to closely integrate obstacle detection into the perception unit.

The paper is organized as follows: First we will shortly discuss the state of the art regarding robot systems designed for manipulation tasks. In section III we will briefly present our robot system and the intended application. Section IV will describe in more detail the components that work together in a manipulation task. In section V we focus on the integration of the components presented in section IV into a fully functional manipulation subsystem. In section VI experimental results are presented and we will end with a summary in the last section.

II. STATE OF THE ART

There is a large number of systems known that examine a similar scenario to ours, so we will mention only a few exemplary systems. One of the earlier ones is the Mobman [1], a system equipped with a single arm and a simple gripper that was presented in 2002 and was able to grasp simple objects that needed to be located on a table.

The ARMAR system [2] is a two armed humanoid with manipulative, perceptive and communicative skills. The focus of the robot Justin by DLR [3] lies on the dexterous two-handed manipulation and control methods for the arms and hands.

Khatib and Brock [4][5] put a strong focus on mobile manipulation under hard dynamic constraints and planning in such dynamic scenarios. Other work related to planning of manipulation tasks are considering the manipulation tasks embedded in the interaction loop with humans [6].

Considerable research work has also been focused on the development of humanoid robots [7], [8] which are however mostly only equipped with simple and light gripping devices and have relative poor manipulation skills. Wide public attention has been obtained by the humanoids from Toyota, but the technical details of these systems are kept secret.

III. APPLICATION AND SYSTEM DESCRIPTION

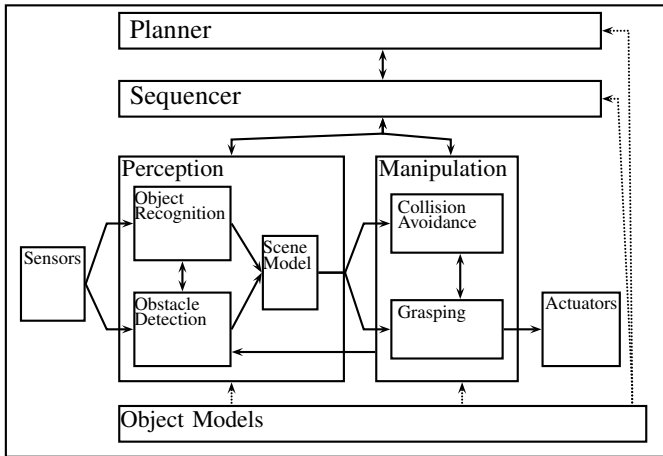


Fig. 2. System overview DESIRE Manipulation modules

The project DESIRE is developing a mobile service robot that executes typical household tasks such as to perform bringing services or to clean up a room. Since it operates in an everyday environment, it can recognize and track people as well as communicate with them. The experimental platform of DESIRE consists of an omni-directional drive, a body, one sensor head and two manipulators (Fig. 1). The sensor head is mounted on a pan-tilt-unit and consists of a synchronized stereo camera system (AVT PIKE 145C) and a 3D time-of-flight (TOF) camera (SwissRanger SR3000) (Fig. 3). Each manipulator consists of an anthropomorphic hand (SAH) mounted on a seven degree of freedom (DOF) arm (KUKA LWR



Fig. 3. Sensor head: Stereo cameras and 3D-TOF camera mounted on a pan/tilt unit

3). The architecture of the components, that interact during manipulation is depicted in Fig. 2.

In a complex system with various subsystems each performing its individual task, the integration of all separate functionalities to form a whole has to bridge the borders between hardware and software. Therefore, the DESIRE project uses a custom-designed CORBA-based middleware. Each component offers its functionality through an interface formulated in the interface definition language (IDL). Since the components are strictly separated from each other, they can easily be integrated into other platforms as well.

When different components work together in order to accomplish a common task, it is crucial that the models they use are consistent throughout the system. Therefore, in DESIRE, all object models are based on a single raw data set that is generated using the Interactive Object Modeling system described in [9]. Each object considered constitutes a raw data entity with a 3D triangle mesh of its surface and 396 stereo images taken from different view points. Of course, the 3D triangle mesh and the stereo views are registered with respect to each other, so that a unique object coordinate reference frame is established.

IV. COMPONENTS FOR COLLISION FREE MANIPULATION

First we will describe the main components that are contributing to the manipulation (Fig. 2) in more details, before we focus on the integration and interaction aspects of those in section V.

A. Object Localization

To achieve the high precision in full 6D object localization that is essential for robust robotic grasping with a multi fingered hand as done by Xue et al. [10] we decided to rely on a stereo camera setup in combination with sift features, a state of the art object recognition method [11] based on local features.

The complete sensor equipment of the robot has been described in section III, for the use in the manipulation it is essential to precisely calibrate the stereo camera system. Both, intrinsic and extrinsic calibration, of the cameras are carried out with the Camera Calibration Toolbox [12] for MATLAB, using 60 stereo image pairs.

As a preprocessing step we construct a 3D sift model of each object, leading to a different type of interest point. The normal

sift interest point consists of a 2D location, a scale, an orientation and a descriptor $\{u, v, s, \phi, d_i\}$ whereas the 3D interest point consists of a 3D location, a scale, a viewpoint direction, an orientation and a descriptor $\{x, y, z, s, x', y', z', \phi, d_i\}$.

The recognition and localization process consists of the following steps:

- 1) Calculating standard sift features in of the images
- 2) Finding hypothesis with high consistent support from these sift features: object classification
- 3) Calculate sift on other image, find stereo correspondences using the descriptor and the epipolar constraint and construct 3D sift features
- 4) Solve the 3D dot cloud matching with ICP: object localization
- 5) Transform the resulting 6D pose into the world coordinate system using the robot's pose and the pan tilt unit's configuration



Fig. 4. Recognition result: 3D dot clouds projected onto left camera image

After having calculated the first set of sift features that commonly consists of about 1.500 - 8.000 features, the correspondences with reference to the robots object database have to be established. We use a kd-tree to perform the matching as the number of interest points reaches high numbers with a growing number of objects. A database consisting of only eight objects for instance already consists of about 1.470.000 features.

The nature of the sift algorithm determines that a classification already comes with a rough localization, therefore we will talk about hypothesis, meaning a classification and a localization. Each sift point, that is matched to a sift point in the database already constitutes such a full hypothesis, when the camera position from which the database interest point was taken is known. This means, that a single image of a scene consists of about 2500 single hypothesis. The problem of classification resolves to the search for local peaks in the seven dimensional hypothesis space. Each peak that is found marks an area with high support by interest points.

After having found a peak the classification problem is solved so the next step it to localize the object with reference to the class. Using the standard method for pose estimation based on only one picture, the POSIT [13] algorithm, we were not able to fulfill our precision goals, so we decided to use a stereo approach that is naturally suited to increase particularly the low precision of the distance measurement which is the main drawback of POSIT.

Our approach takes all 2D sift points of a peak and searches for the corresponding features in the other camera's image.

For each interest point that was found in both pictures a 3D position in the camera frame can be determined using standard triangulation methods, given a precise stereo calibration.

The resulting 3D dot cloud is subset of the object's 3D sift model in the database with known correspondences, so a simple iterative closest point method delivers the 6D pose of the object in the camera frame. To account for mismatched interest points the result of the ICP is evaluated regarding the distance of the interest point pairs, dropping the ones that have a higher distance than twice the mean distance. If this process drops pairs the ICP is evaluated again. In the end a theoretical minimum of three pairs is required to find a pose, however experiments showed that high accuracy requires a minimum of six pairs.

One important advantage of this 3D sift stereo approach is, that in contrast to other approaches as presented in [14] it is able to handle objects of any shape, as long as the object fulfills the requirements on it's texture that are inherent to the sift algorithm. In a further stage it is planned to incorporate state of the art methods for non-textured object recognition to increase the range of object types the robot can handle.

B. Obstacle Detection

For a collision free path and grasp planning a full model of the scene is necessary. Since the object localization is only able to detect known objects an obstacle detection is required to complete the scene model.

In our system, we rely on a 3D time-of-flight camera that has the advantage that it delivers 3D information at a high frame rate. Since time-of-flight cameras are still somewhat prototypical, there are numerous factors that affect the accuracy of the distance measurements. A lot of research is being done to understand the effects and to compensate the errors introduced [15], [16], [17]. However, a complete model to correct all of them is still missing.

In order to get rid of the distortion of the lens and the misalignment of the chip, the sensor is calibrated using the MATLAB Camera Calibration Toolbox [12] based on 100 intensity images. By means of the estimated intrinsic and extrinsic camera parameters, a point cloud in Cartesian coordinates corresponding to the radial distance measurements of the depth map can be computed.

The distance measurements of time-of-flight cameras experience considerable noise. Therefore, smoothing of the data is essential. Furthermore, jump edges (i.e. gaps in distance between neighboring pixels due to occlusion) are integrated over such that the gaps are filled with pseudo measurements. We process the sensor's depth map with a combination of edge-preserving smoothing and jump edge-finding in order to filter the pixels corresponding to pseudo measurements (Fig 5).

The 3D points that persist preprocessing form the data basis for the obstacle representation. In our approach we coarsen the detailed 3D point cloud with bounding volumes (such as boxes, cylinders, etc.) as an approximation. In a further stage, we intend to approximate the obstacles by triangle meshes in order to improve the systems accuracy.

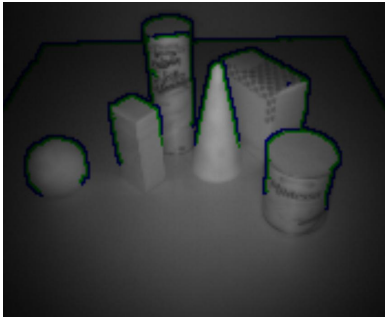


Fig. 5. Jump edges of the objects

C. Collision Free Grasping

The grasping simulator “GraspIt!” [18] is utilized to plan high quality grasps. It uses hand preshapes and automatically generated grasp starting positions and directions. The method is based on the observation that humans unconsciously simplify the grasping task to selecting one of only a few different preshapes appropriate for the object and for the task to be performed. The surface object model in the introduced object database is decomposed into a small set of primitive shapes such as spheres, cylinders, cones and boxes. Starting positions and approach directions are generated from these primitive shapes. In the simulation, the hand in predefined preshape is placed at a starting position, moves along one of the approach directions towards the object and then closes around the object. After the object is grasped, all the contact points are collected and the grasp quality is evaluated. “Largest sphere in grasp wrench space” is used as grasp quality measurement, so that the grasp can resist with independence of the perturbation direction. The found grasps with high grasp quality are saved into a grasp database. During the execution, after the localization of the object by the proposed method with SIFT features, a grasp for the object is searched in the grasp database, which can be performed by the robotic arm and without any collision with the environment.

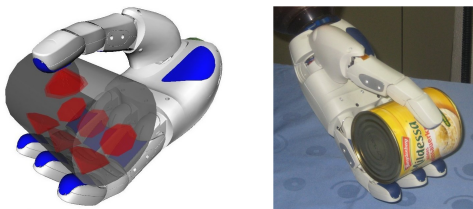


Fig. 6. Simulated and real grasp

Other localized objects and the bounding volumes which the unknown objects approximate are treated as obstacles in the environment. After a feasible grasp is found, where the fingers of the robotic hand only collide with the object at desired contact points and without any other collision between the robot and the environment, a probabilistic collision free path planner [19] is used to bring the robotic arm to the starting position. After the object is grasped, it is also treated as part

of the kinematic chain for the following collision free path planning to avoid collisions between the grasped object in the hand and the obstacles.

V. SYSTEM INTEGRATION

The first prerequisite of an integrated system is a precise common calibration throughout the components. We will describe this process in our system and then show the integration of known object localization and obstacle detection and their interaction that is necessary to achieve a full scene model which allows for manipulation.

A. Calibration

When different components interpret their environment and establish a common scene model for it, it is crucial that they all settle on a common coordinate system. The sensor head is calibrated as described in section IV-A. Furthermore, a hand-eye calibration is carried out in order to determine the transformation between the sensors and the actuators. By means of the calibration, the components utilizing sensors or actuators work in the same coordinate reference frame. We establish a world coordinate system in which all interpretation of the environment is done. This means, each component must transform between its own coordinate system and the one of the world whenever data is provided to other components.

In order to enhance the data quality of the time-of-flight camera, we propose an auxiliary calibration step. However since it requires scene information, it is only feasible in an integrated system such as ours. The sensor usually possesses considerable variability over its depth sensing range. As a consequence, planar surfaces are often displayed vaulted. As described in the literature [15], the effect can be corrected by means of a plane calibration. With this offline correction, planes get transformed to planar point clouds. In the strict sense though, the depth correction is only valid under the conditions present during calibration such as the surface reflectance and the angle of intrusion. Our experiences with the sensor show that though planar, point clouds corresponding to planes still may be shifted and tilted in depth. By means of our online correction, we transform the point cloud in such a way that it fits those parts of the environment best that are assumed to be known. Thereto, planes in the robot’s environment (such as walls, tables, etc.) may be used as references which are identified in the point cloud by plane fits. The deviations between the planes and their fits define depth correction values in each pixel. In order to perform the correction, the obstacle detection uses the integrated scene model as a source for information on the known parts of the environment.

B. Integration of Object Localization and Obstacle Detection into a Full Scene Model

The main task in the integration of known object localization and obstacle detection is to ensure that no known objects are additionally classified as obstacles. Running the obstacle detection stand-alone, this will be the case for every localized object. A comparable situation is induced by the robot itself,

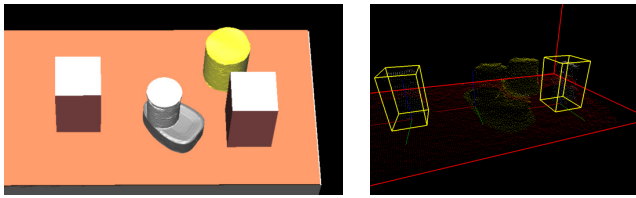


Fig. 7. Scene model (a) and dot cloud (b)

as it's actuators may appear in the field of view where they will consequently be classified as obstacles. In both situations, entities that are already known to the system are falsely classified as obstacles, prohibiting every attempt to execute manipulation tasks. To achieve a full consistent scene model, known objects and the robot itself are reported to the obstacle detection, that way the obstacle detection becomes an unknown object detection that can be consistently joined with the known objects.

After having received the information about known objects and the robot actuators state, the obstacle detection renders an artificial view of the known objects and the actuators in the time-of-flight camera's view point. This artificial measurement is used as a mask on the current measurement to find all points that can not be explained by known objects or the robot. These are classified as unknown objects, approximated by bounding boxes and passed to the scene model.

Additionally the obstacle detection also checks whether or not the space occupied by recognized objects is really taken. In case the space is free, it reports to the scene model which in turn may try to clarify the situation by a recognition attempt. Since this is not contributing to the task of manipulation this aspect is of minor interest here. By means of these methods, we obtain a fully integrated scene model that respects recognized and localized objects as well as detected obstacles consistently.

VI. EVALUATION AND EXPERIMENTAL RESULTS

A. Evaluation of Object Localization

To evaluate the suitability of our system for the manipulation task we start with an evaluation of the object localization accuracy: Therefore we place two objects from our database at 59 different locations within the workspace of the robots hands in front of the sensor system. We compare the resulting 6D pose of our sift based algorithm with ground truth values that were determined using a precise calibration pattern. The results presented in Tab. I are given in the coordinate system of the camera. These results show that the accuracy of the localization is high enough to execute robust grasps.

B. Evaluation of Hand-Eye Chain

Knowing that the accuracy of the localization itself is sufficient for grasping, the whole chain from the cameras to the fingertips has to be evaluated. In the measurement setup we replace the hand with a calibrated probe and measure the

TABLE I
ACCURACY OF THE LOCALIZATION IN THE AREA OF INTEREST

Dimension	stddev/mm	mean error/mm
x	3.54	-1.79
y	1.86	3.54
z	2.87	4.65



Fig. 8. The measurement setup

position of a calibration pattern that can be localized very accurately with this probe (Fig. 8).

TABLE II
ACCURACY OF THE HAND-EYE CHAIN

Dimension	stddev/mm	mean error/mm
x	3.1004	1.5512
y	1.4423	-3.469
z	1.6929	0.9576

The results presented in Table II originate from a set of 53 measurement points that are located throughout the working volume of the robot. Since KUKA reports the absolute positioning accuracy and repeatability of the LWR to be below one millimeter, the small errors measured here originate from the calibration in the camera, the pan-tilt unit and the mounting point of the arm. The accuracy measured here is high enough to allow secure grasping with the previously describes setup.

C. Evaluation of Manipulation Task

To evaluate the suitability of our system for manipulation operations, we utilize a part of a "cleanup table" scenario. In this scenario, multiple objects, known and unknown, are arranged on a table in front of the robot, possibly occluding each other.

The grasp planning has to find an accessible, collision free grasp, as well as collision free approach and the deproach paths to be able to accomplish a full collision free manipulation.

In comparison to our previous experiments the manipulation task introduces new sources of errors. The hand was not included in the hand-eye-calibration test since it's real TCP



Fig. 9. A successful manipulation

is difficult to measure, so impreciseness in the hand model might influence the result of the experiment. Another problem is introduced by the motion of the fingers when closing a grasp. It is not guaranteed, that all fingers reach their target pose at the same moment, especially, since impedance control is used in the final phase of grasping. A faster finger can displace the object, although it is correctly located. Due to this fact, object displacement is ruled out as a quantitative result.

It is difficult to define a metric on this type of experiment, thus we evaluate the success of the manipulation. In a successful manipulation, the robot must grasp the object in the way that is intended by the grasp planning. It must lift the object and may not drop it again. No collisions may occur, neither between the hand and an object nor between the manipulated object and another object. The robot may move the object during the closing of the grasp. On the other hand, if objects are placed too dense, it might not be possible for the robot to find a collision free grasp at all. If this is detected and reported by the system correctly, the run does not count as a failure. Over the period of three weeks we performed many manipulations and noticed that no collision with obstacles ever happened, but due to over-secureness of the bounding boxes it often occurred, that no grasp was found.

In the end we performed another test of 20 manipulations in different scenes, where object were partly occluded but not too close to the target object to account for the over secure bounding boxes. All of these test were executed successfully, so it seems that the localization accuracy achieved by the system is high enough to grasp successfully.

VII. CONCLUSION

The DESIRE robot is able to accomplish manipulation tasks in everyday life environments, as it is able to acquire a complete model of the scene by integrating known object localization and obstacle detection. Experiments showed, that the calibration and the accuracy of all components is high enough to allow collision free manipulation tasks.

ACKNOWLEDGMENTS

This work was partly funded as part of the research project DESIRE by the German Federal Ministry of Education and

Research (BMBF) under grant no. 01IME01D, 01IME01E and 01IME01A.

REFERENCES

- [1] G. v. Wichert, C. Klimowicz, W. Neubauer, T. Woesch, G. Lawitzky, R. Caspari, H.-J. Heger, P. Witschel, U. Handmann, and M. Rinne, "The robotic bar - an integrated demonstration of man-robot interaction in a service scenario," in *Proceedings of the 2002 IEEE Int. Workshop ROMAN, Berlin*, 2002.
- [2] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstien, A. Bierbaum, K. Welke, J. Schroeder, and R. Dillmann, "Toward humanoid manipulation in human-centred environments," *Robotics and Autonomous Systems*, vol. 56, pp. 54–65, 2008.
- [3] C. Ott, O. Eiberger, W. Friedl, B. Baeuml, U. Hillenbrand, C. Borst, A. Albu-Schaeffer, B. Brunner, H. Hirschi, S. Kielhoefer, R. Konietschke, M. Suppa, T. Wimboeck, F. Zacharias, , and G. Hirzinger, "A humanoid two-arm system for dexterous manipulation," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, Genova, Italy, 2006, pp. 276–283.
- [4] O. Khatib, O. Brock, K. Chang, D. Ruspini, L. Sentis, and S. Viji, "Human-centered robotics and interactive haptic simulation," *International Journal of Robotics Research*, vol. 23(2), pp. 167–178, 2004.
- [5] Y. Yang and O. Brock, "Elastic roadmaps: Globally task-consistent motion for autonomous mobile manipulation," in *Robotics: Science and Systems*, Philadelphia, USA, 2006.
- [6] E. Sisbot, L. Marin-Urias, R. Alami, and T. Simeon, "A human aware mobile robot motion planner," *IEEE Transactions on Robotics*, vol. 23 (5), pp. 874–883, 2007.
- [7] K. Akachi, K. Kaneko, N. Kanehira, S. Ota, G. Miyamori, M. Hirata, S. Kajita, and F. Kanehiro, "Development of humanoid robot hrp-3p," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, 2005, pp. 50–55.
- [8] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura, "The intelligent asimo: system overview and integration," in *IEEE/RSJ International Conference on Intelligent Robots and System*, vol. 3, 2002, pp. 2478–2483.
- [9] P. Becher, R. Steinhaus, R. Zoellner, and R. Dillmann, "Design and implementation of an interactive object modelling system," in *Proceedings of the Joint Conference on Robotics (ISR 2006 ROBOTIK 2006)*, Munich, Germany, 2006.
- [10] Z. Xue, J. Marius Zoellner, and R. Dillmann, "Grasp planning: Find the contact points," in *IEEE International Conference on Robotics and Biomimetics*, 2007.
- [11] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1150–1157.
- [12] J.-Y. Bouguet, "Matlab calibration toolbox." [Online]. Available: "http://www.vision.caltech.edu/bouguetj/calib_doc/"
- [13] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision*, Springer Netherlands, vol. Volume 15, Numbers 1-2, pp. 123–141, 1995.
- [14] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6d object localization for grasping with humanoid robot systems," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, San Diego, CA, USA, 2007.
- [15] I. H. Kahlmann T., Remondino F., "Calibration for increased accuracy of the range imaging camera swissranger," 2006.
- [16] M. Plaue, "Analysis of the pmd imaging system," University of Heidelberg, Germany, Tech. Rep., 2006.
- [17] S. Guomundsson, H. Aanaes, and R. Larsen, "Environmental effects on measurement uncertainties of time-of-flight cameras," *Signals, Circuits and Systems, 2007. ISSCS 2007. International Symposium on*, vol. 1, pp. 1–4, July 2007.
- [18] A. Miller and P. Allen, "Grasplit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.
- [19] M. Saha, G. Sanchez, and J. Latombe, "Planning multi-goal tours for robot arms," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, Taipei, Taiwan, 2003.