Natural Demonstration of Manipulation Skills for Multimodal Interactive Robots

Markus Hüser, Tim Baier-Löwenstein, Marina Svagusa, and Jianwei Zhang

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department Informatics, Group TAMS
{hueser,baier-loewenstein,9svagusa,zhang}@informatik.uni-hamburg.de

Abstract. This paper presents a novel approach to natural demonstration of manipulation skills for multimodal interactive robots. The main focus is on the natural demonstration of manipulation skills, especially grasping skills. In order to teach grasping skills to a multimodal interactive robot, a human instructor makes use of natural spoken language and grasping actions demonstrated to the robot. The proposed approach emphasizes on four different aspects of learning by demonstration: First, the dialog system for processing natural speech is considered. Second, an object detection and classification scheme for the robot is shown. Third, the correspondence problem is addressed by an algorithm for visual tracking of the demonstrator's hands in real time and the transformation of the tracking results into an approach trajectory for a robotic arm. The fourth aspect addresses the fine-tuning of the robot's hand configuration for each grasp. It introduces a criterion to evaluate a grasp for stability and possible reuse of a grasped object. The approach produces stable grasps and is applied and evaluated on a multimodal service robot.

1 Introduction

The field of human-robot interaction has seen a lot of advances in recent years. Especially, the way how robots can be programmed by simple teaching-by-showing via teach pendants, explicit robot programming languages, or task-level programming languages. The latter language enables the user to give sub-goals to the robot, rather then defining each single action the robot has to perform to solve a given problem. Although robot programming has been simplified by these developments, only a few and skilled experts are able to handle them. Robot programming still lacks the benefits of human teaching and learning such as a natural and versatile learning scheme of demonstration and imitation. In this paper we propose a framework for natural demonstration of manipulation skills, especially grasping skills. These skills are basic for a service robot and crucial to manipulation tasks and interaction with the environment. Service robots operate in scenarios which demand very flexible and robust behaviors like e.g. office environments. This robustness to non-static environments cannot be achieved by the classic robot programming techniques. Therefore, it is important to further simplify interaction between robots and humans.

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2007, LNCS 4555, pp. 888–897, 2007.

The proposed framework meets these challenges through the integration of multi-modal techniques for human-robot-interaction and enables a user to program a service robot by the scheme of learning by imitation. The natural interaction between the user and the robot is emphasized by only using spoken dialog and gestures performed by the user. Neither additional sensors like data gloves, which have to be worn by the user, nor complicated sensor setups installed in the environment are needed.

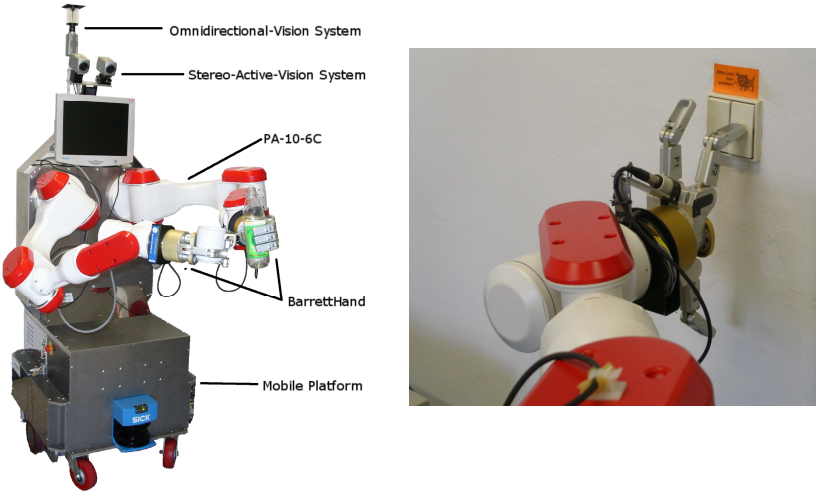

Fig. 1. TASER – TAMS Service robot (left), TASER operating a light switch (right)

The proposed approach is applied to the service robot of the TAMS Institute at the University of Hamburg, the **TA**MS-**Se**rvice-**R**obot (TASER). Figure 1 shows the robot TASER and how it is manipulating a light switch. The robot operates during normal workdays in an office environment. It is used as an experimental platform to point out the usability of the proposed approach.

The remainder of the paper is organized as follows: In section 2 an overview of related research is given. Section 3 covers natural demonstration of manipulation skills, whereas in section 3.1 the service robot TASER and the scenario is introduced. Section 3.2 describes the dialog system. Section 3.3 focuses on learning by demonstration and section 3.4 examines grasp analysis. Conclusions and future work are presented in Section 4.

2 Related Research

A good overview of learning by demonstration is given in [1]. A multimodal approach using language, vision and motor based on a hierarchical architecture is presented in [2]. It enables a student robot to learn three different behaviors (‘go’, ‘pick’, ‘lift’) from a teacher robot. A model for robot imitation also using language, vision and motor input based on the concepts of the modularity in the brain and mirror neuron system is given in [3].

To determine the best imitation strategy, a method for extracting goals and constraints of a demonstrated task is presented in [4]. Simple reaching tasks are considered and captured via magnetic field sensors and stereo-cameras and encoded into Hidden Markov Models (HMMs). Trajectories are reproduced according to a metric based on invariants of time, which measures the quality of reproduction.

The work of [5] is built on the closed-world assumption, which emphasizes the exact observation of the human operator. The whole work cell is modeled using accurate scene data acquired by data gloves and stereo cameras. In the work cell dual-arm manipulation tasks are demonstrated for humanoid robots [6]. For the reason of robust and accurate data acquisition (as shown in [4], [5], [6]), data gloves and magnetic sensors are common tools to observe the demonstrator's actions. But techniques for contact-free and easy-to-use human-robot-interaction interfaces are important to simplify the application of learning by demonstration. Especially the imitation of manipulation tasks requires a robust tracking of the human hands. There are many approaches to this topic, e.g. [7] proposes a method for the tracking of articulated finger motion in monocular images, where a geometric model of a hand is used to generate contours which are matched with observed monocular edge and skin color images. A hierarchical Bayesian filter is developed to allow the integration of temporal information.

In [8] a method for imitation learning based on visuo-somatic mapping, ranging from the observed demonstrator's posture to remembering the self-posture, is presented. This happens via mapping from the self-motion observation to the self-posture of a robot. The mapping from postures to posture space and the mapping from trajectories in posture space onto a motion-segment space is both done by SOMs. Optical flow from the demonstrator's motion is mapped onto a flow-segment space where flow data is connected with the corresponding motion segments in motion-segment space. The connection with self-motion is done via Hebbian Learning.

A brief overview of the special field of robot grasping is given in [9]. Canny et. al. [10] presents a grasp quality measure based on the approximation of the grasp wrench space (GWS). Due to the computational complexity of the problem Borst et. al. [11] presents a method to reduce the complexity by using a specialized task wrench space for object grasping. In [12] the author presents a method to generate grasps from prototypes using an object wrench space taking the object geometry into account. An efficient way to check form closure properties of a grasp is given in [13] where the authors present an approach to find form-closure grasps for n-hard fingers of 3-D objects, which are represented by a model of discrete points. In contrast Han et al. [14] suggest a method transforming the problem into linear matrix inequalities (LMIs), which solves grasp evaluation and grasp force optimization.

Most of the presented approaches above use the force or form-closure of a grasp wrench space as criterion for a grasp. This criterion guarantees resistance of the grasp against any external force, which might be applied to the grasped object. This often leads to large forces and torques at the grasping points. [15] considers this problem and proposes a method where a task wrench is generated and mapped to the grasp matrix in the form of a virtual contact acting with a certain magnitude. By the use of a LMI method it is possible to find the force necessary to resist the internal force.

3 Natural Demonstration of Manipulation Skills

In the following the approach to natural demonstration of manipulation skills and its implementation on the service robot TASER (section 3.1) is presented. The main focus is on the natural demonstration of grasping skills, which has to produce stable grasps and consider the reuse of the grasped object. For an easy and naturally-to-use application of this technique it is important to embed the learning process into a versatile interface for human-robot-interaction, which is capable of a dialog-driven communication between the robot and its user. The developed dialog system is intended to make the interaction between the robot and humans simpler and more intuitive and therefore uses speech-recognition alongside with speech-generation to drive the dialog (see section 3.2).

However natural demonstration also requires the robot to capture the data by itself without restricting the environment by the demands of special hardware, like data gloves, magnetic field sensors, head-mounted eye trackers or special sensor installations, which are cumbersome to set-up, calibrate and use. Therefore, only the onboard sensors (video, audio, laser) of the service robot TASER are used to gather data, when a human teacher demonstrates a grasping task. In doing so, it is important that the robot first detects an action and learns how to reproduce this action and then decides which type of grasp is suitable for further handling of the grasped object considering different constraints, e.g. like the fill level of a cup of tea. In the case of grasping a specific object, the robot has to detect the object (section 3.3), learn an approach trajectory and the tag point (section 3.4) as well as an adequate force for a stable grasp and the type of grasp with respect to the further handling of the grasped object (section 3.5).

3.1 The Service Robot TASER in Its Office Environment

A lot of service robots described in literature are built from special hardware which is designed for certain robot systems only. Furthermore, their manipulation and grasping skills seem to be rather limited because many robots only have a two-finger parallel jaw gripper. One of the particular objectives in building the robot TASER [16] was to assemble it mainly from off-the-shelf hardware.

TASER's mobile platform is equipped with a differential drive with integrated wheel encoders, two laser-range finders and a gyroscope for navigation. The localization reaches an accuracy of ±1cm in position and ±1° in orientation. The manipulator system of TASER currently consists of two Mitsubishi PA10-6C robot arms. A BH-262 BarrettHand and a micro-head camera are attached as a tool to each arm. With this design the system attains a humanlike workspace and silhouette. The PA10-6C has six degrees of freedom (DoF) and a kinematic length, which is similar to the length of a human arm. The BarretHand is a three-finger robot hand with 8 DoF, force sensors and a TorqueSwitch mechanism similar to a human hand [17].

TASER's interface for human-robot-interaction is intended to provide a simple and intuitive interaction between the robot and humans. The main sensors for the interaction are several camera systems mounted on the robot. In addition to the micro-head cameras on the hands, the system has a stereo-camera system mounted on a pan-tilt unit and an omnidirectional vision system (see Fig. 1). In order to teach TASER a

grasping skill, a human instructor uses natural speech and demonstrates the grasping action to the robot. This involves a human instructor standing opposite the service robot (see Fig. 2) and demonstrating a grasping skill by saying 'start', performing the grasp and then saying 'stop'. This process is repeated several times until the robot has collected sufficient data to analyse according to forces and reusability.

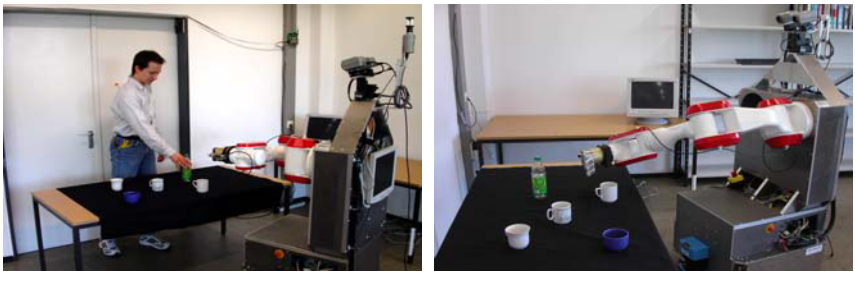

Fig. 2. Instructor demonstrates a grasping skill (left) and TASER reproduces the skill (right)

3.2 Dialog System

A finite-state based dialog system was developed for the robot. The robot can be queried for information or given a task through the dialog-system. It is a model of a finite-state automate, which contains a sequence of predetermined states [18] and is used to guide the user through a dialog. A rough overview of the dialog system is given in Fig. 3. Every time a dialog step is completed, which means that all the required information has been given by the user, the system state changes and reaches its end state. The advantage of a finite-state based system compared to a frame-based system, which has no predetermined dialog, or compared to an agent-based system, which includes intelligent behavior, is its simplicity. The semantics of the system is clear and intuitive. Moreover, as the user's responses are restricted, fewer technological demands are put on the system components, particularly the speech recognizer. Additionally, there is a smaller likelihood of errors in the more constrained system. An example of a dialog in which the system verifies the user's input [19] is given in table 1.

The instruction for grasping an object is received by speech recognition. The recognized sentence is passed to the verifier, which parses the instruction. The instruction is transferred to the speech output, which repeats it. If the user confirms the speech output to the dialog system, this confirmation is brought to the analyzer, which consists of the dialog manager, the context manager and the controller. If the user does not confirm the speech output of the robot, the dialog-system returns to the subsequent state of the last confirmed state. The user can then repeat his speech input. Thus, only instructions confirmed, are processed by the analyzer. Each command is assigned to a context and administrated by the context manager. The command ['Grab', 'Cup'] consists of a list, which defines current and further possible functions and objects in this context. In the example above, the possibility of two

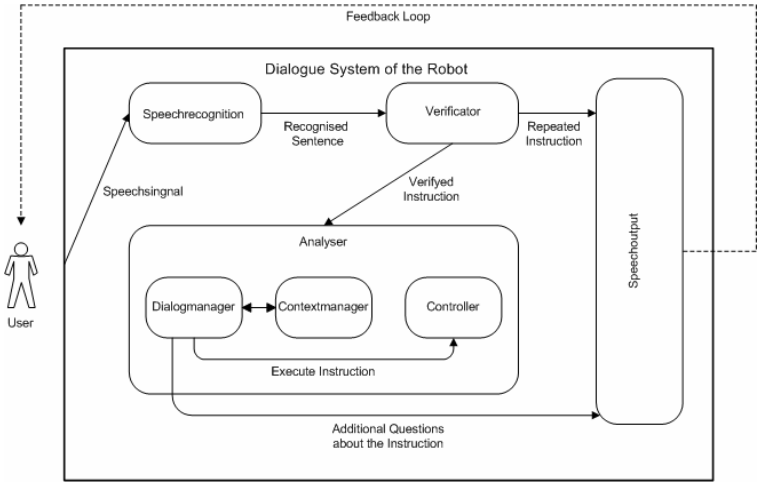

Fig. 3. Dialog system

identical cups, which can both be grasped is given. This constitutes a mnemonic ambiguity [20], which takes place when an input reference is ambiguous to the robot's context. The dialog manager gets the list from the context manager and initiates a speech output, which dissolves the ambiguous user instruction. After all these steps are accomplished, an order is sent by the dialog manager to the controller to execute the instruction.

Table 1. Resolved ambiguity in a dialog between user and robot

User	Robot
Get the cup.	
	May I bring you the cup?
Yes	
	There are two cups, would you like the green cup or the red cup?
The green cup	
	Do you want the red cup?
No	
	There are two cups, would you like the green cup or the red cup?
The green cup	
	Do you want the green cup?
Yes	
--- Robot is grasping the green cup ---	
	Done. Can I do something else for you?

3.3 Learning by Demonstration

Learning by demonstration is used to determine the trajectory of the instructor's hand while performing the grasping action and to estimate the tag point, where the hand touches the object. But before the robot can do so, it has to detect the object, its position and its orientation with the help of its stereovision system. This is done by

the use of scale-invariant features (SIFT) [20]. To each object several SIFT key points are assigned and correspondence is established between the stereoscopic images. Therefore, the coordinates of those key points can be measured in 3D. The position and orientation of the object is determined by solving a linear equation system.

To learn a grasping skill through demonstration, it is necessary to know how the demonstrator grasps an object. Therefore the trajectory as well as the configuration of the hand, e.g. the articulation of the fingers, has to be known. In order to track the demonstrator's hands in stereoscopic images in real-time, active contours [21] and an expectation-maximization-like algorithm [22] is adapted and applied to local binary patterns [23] and color histograms [24]. The robot repetitively tracks the demonstrator's grasping actions and feeds the tracking results into a three-dimensional self-organizing map [24], to minimize discrepancies between intended and tracked trajectories. In this process the robot's dialog system is used to inform the robot about the time when the instructor will start and end his demonstration. The topology of the SOM is arranged to correspond to the three-dimensional space in which the user's hand is tracked. After convergence, the SOM gives a spatial description of the collected data and serves as the input data structure for an RL algorithm. The RL algorithm finds trajectories and hand configurations optimized for use by the robot arm and hand (Fig. 4) and serves as starting point for the following grasp analysis.

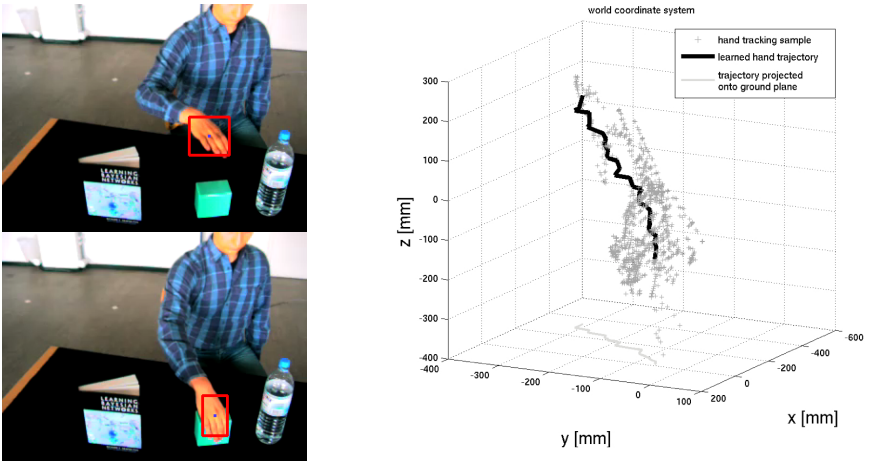

Fig. 4. Hand tracking results and generated trajectory of box grasping skill

3.4 Grasp Analysis

Most of the grasp evaluation criteria presented in section 2 are related to forces and torques. This is very useful if stability is the main criterion in grasping objects with an artificial hand. Our approach goes beyond this and defines a criterion which is related to the reusability of an object grasped with a certain grasp. The criterion has been introduced in [25] and [26]. For example the grasp shown in Fig. 5 might be very stable according to forces and torques, but it is unusable if somebody wants to fill something into the cup. For this reason we defined four sub-criteria related to

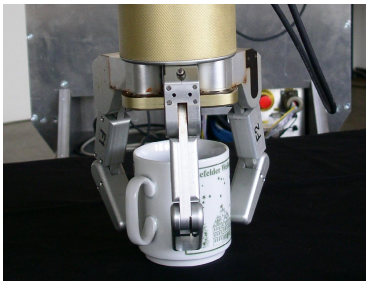

Fig. 5. Grasping a mug with a BarrettHand

operations being performed with a grasped object that is evaluated automatically during grasp analysis. A rating for the grasp is generated from the evaluation result. For each criterion, a list of constraints is defined which must be satisfied.

One of the four criteria is the operation *Pour-In.* It is defined as a gain in weight and a free area around the feed opening of the object. The free space around the ``feed opening" of e.g. the cup is defined as “keep-out-area” with the shape of a flat cylinder around the top of the object. The weight gain could be formulated in the form of a task wrench space, like in [12] or [15].

Similar to the Pour-In operation the operation *Pour-Out* can be defined as a loss in weight and a free area around the “feed-opening”. The keep-out-area makes sure that the opening is nod occluded by parts of the manipulator. Additionally, the grasp has to resist the forces which occur during the rotation movement.

The main criterion for a service robot is the *Handover* operation. It is defined by the free spaces on the object, which are needed to pass the object from one hand to the

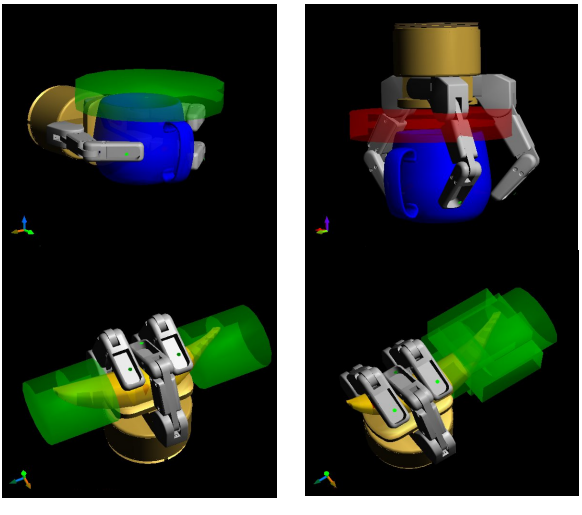

Fig. 6. Simulation of grasping a mug and evaluation of the Pour-In criterion (top row), evaluation of the Handover criterion for grasping a banana (bottom row)

other. The space is firstly defined by the shape of the object and secondly by the position which the opponent wants to grasp. Examples for handover grasp evaluation are shown in Fig. 6 (green volume marks spaces for possible Handover operations).

For some objects it is useful to define the operation *Movement* as a criterion for a grasp. For the evaluation of the criterion, the forces occurring during execution of a user-defined motion are considered. The motion can be defined via a transform or as a combination of basic movements like up, down, left, right, rotate etc. After the constrains for the selected criteria have been checked, a score for the object is built from a weighted sum of the four computed grasp criteria where s is the resulting score a_j is the factor of each constraint and C_j is the value for the constraint.

$$s = \sum_{j=1}^{n} a_j \; C_j \tag{1}$$

If the user restricted the object's purpose to some of the criteria the weighted sum can yield different results in different contexts. E.g. it is useless to compute if Pour-Out and Pour-In operations are possible for a hammer.

4 Conclusions

The realization of a framework for natural demonstration of grasping skills on a multimodal interactive service robot has been presented. It enables the user to perform a natural human-like learning process with the service robot TASER. A dialog system based on natural spoken language makes the interaction between the robot and humans simpler and more intuitive. Furthermore, it is used to solve problems of ambiguity and keeps the users hands free while demonstrating a grasping action. Additionally the dialog system can be used to command the robot e.g. to carry out tasks like turning the light on or transporting items.

The criteria used for grasp evaluation have the great advantage that they are easy to understand, even if somebody has no idea about robotics or physics. They further simplify the interaction and communication process between the robot and a human instructor. They also enable the robot to choose task-specific grasps from a grasp database according to the current context.

The next step will be to extend the framework to increase TASER's usability. In this step the variety of the high-level functions like object grasping and multimodal interaction will be increased by more distinct manipulation of objects, like taking a printer's output and transport it to the user.

References

1. Billard, A., Siegwart, R.: Robot Learning from Demonstration. Robotics and Autonomous Systems, vol. 47(2 & 3) (June 30, 2004)
2. Weber, C., Elshaw, M., Zochios, A., Wermter, S.: A Multimodal Hierarchical Approach to Robot Learning by Imitation. In: Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, pp. 131–134. Genoa, Italy (2004)
3. Elshaw, M., Weber, C., Zochios, A., Wermter, S.: An Associator Network Approach to Robot Learning by Imitation through Vision, Motor Control and Language. In: Proc. of the Int. Joint Conference on Neural Networks, pp. 591–596. Budapest, Hungary (2004)

4. Calinon, S., Guenter, F., Billard, A.: Goal-Directed Imitation in a Humanoid Robot. ICRA, Barcelona (2005)
5. Ehrenmann, M., Zöllner, R., Rogalla, O., Vacek, S., Dillmann, R.: Observation in Programming by Demonstration: Training and Execution Environment. In: Third IEEE International Conference on Humanoid Robots, Karlsruhe, Germany (2003)
6. Zöllner, R., Asfour, T., Dillmann, R.: Programming by Demonstration: Dual-Arm Manipulation Tasks for Humanoid Robots. In: IEEE Int. Conf. IROS, pp. 479–488, Japan (2004)
7. Stenger, B.: DPhil Thesis: Model-Based Hand Tracking Using A Hierarchical Bayesian Filter. Department of Engineering, University of Cambridge, (March 2004)
8. Asada, M., Ogino, M., Matsuyama, S., Ooga, J.: Imitation Learning Based on Visuo-Somatic Mapping. In: Proceedings 9th Int. Symposium on Experimental Robotics (2004)
9. Bicchi, A., Kumar, V.: Robotic grasping and contact: a review. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 348–353 (2000)
10. Ferrari, C., Canny, J.: Planning Optimal Grasps. In: Proceedings of the IEEE International conference on Robotics and Automation, pp. 2290–2295. Nice, France (1992)
11. Borst, C., Fischer, M., Hirzinger, G.: Grasp Planning: How to Choose a Suitable Task Wrench Space. In: Proceedings International Conference on Robotics and Automation, pp. 319–325. New Orleans, LA, USA (2004)
12. Pollard, N.S.: Parallel Algorithms for Synthesis of Whole-Hand Grasps. In: Proceedings of the IEEE International Conference on Robotics and Automation, Albuquerque, NM (1997)
13. Lui, Y., Lam, M.: Searching 3-D Form Closure Grasps in Discrete Domain. In: Proceedings IEEE Int. Conference on Intelligent Robots and System, Las Vegas, Nevada (2003)
14. Han, L., Trinkle, J., Li, Z.X.: Grasp Analysis as Linear matrix Inequality Problems. IEEE Transactions on Robotics and Automation 16(6), 663–674 (2000)
15. Haschke, R., Steil, J., Steuwer, I., Ritter, H.: Task-Oriented Quality Measures for Dextrous Grasping. In: IEEE Conf. on Computational Intelligence in Robotics and Automation (2005)
16. Westhoff, D., Baier, T., Hüser, M., Zhang, J.: A flexible software architecture for multi-modal service robots, Multiconf. on Comput. Engineering in Systems Applications (2006)
17. Kraiss, K.F.: Advanced Man Machine Interaction, Fundamentals and Implementation, pp. 153–162 (2006)
18. McTear, M.F.: Spoken Dialogue Technology, Enabling the Conversational User Interface, ACM Computing Surveys, vol. 34(1), pp. 90–169 (March 2002)
19. Kulyukin, V.: Human Robot Interaction through Gesture Free Spoken Dialogue in Autonomous Robots, vol. 16, pp. 239–257
20. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
21. Hüser, M., Baier, T., Zhang, J.: Imitation Learning of grasping skills through a multimodal service robot, 2005 IEEE ICRA Workshop on The Social Mechanisms of Robot Programming by Demonstration. Barcelona, Spain (2005)
22. Zivkovic, Z., Kröse, B.: An EM-like algorithm for color-histogram-based object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)
23. Mäenpää, T., Pietikäinen, M.: Texture analysis with local binary patterns. In: Chen, C.H., Wang, P.S.P.: Handbook of Pattern Recognition & Computer Vision, 3rd
24. Hüser, M., Baier, T., Zhang, J.: Learning of demonstrated Grasping Skills by stereoscoopic tracking of human hand configuration, 2006 IEEE ICRA, Orlando, USA (May 15-19, 2006)
25. Hüser, M., Baier, T., Westhoff, D., Zhang, J.: Multimodal learning of demonstrated grasping skills for flexibly handling grasped objects, ISR/Robotik (2006)
26. Baier, T., Zhang, J.: Reusability-based Semantics for Grasp Evaluation in Context of Service Robotics. In: IEEE ROBIO 2006, Kunming, China (2006)