# ToBI - Team of Bielefeld: The Human-Robot Interaction System for RoboCup@Home 2014

Leon Ziegler, Jens Wittrowski,
Sebastian Meyer zu Borgsen, Sven Wachsmuth

Faculty of Technology, Bielefeld University,
Universitätstraße 25, 33615 Bielefeld, Germany

**Abstract.** The Team of Bielefeld (ToBI) has been founded in 2009. The RoboCup activities are embedded in a long-term research history towards human-robot interaction with laypersons in regular home environments. The RoboCup@Home competition is an important benchmark and milestone for this goal in terms of robot capabilities as well as the system engineering approach. For RoboCup 2014, we mainly improved abilities for the perception-based understanding of the robot's environment. An Articulated Scene Model (ASM) is used to systematically fuse scene change events with the perception of the 3D room structure. This is further elaborated by an Implicit Shape Model (ISM) for furniture recognition.

## 1 Introduction

The RoboCup@Home competition aims at bringing robotic platforms to use in realistic home scenarios. Thus, the robot needs to deal with unprepared domestic environments, perform autonomously in them and interact with laypersons.

Todays robotic systems obtain a big part of their abilities through the combination of different software components from different research areas. To be able to communicate with humans and interact with the environment, robots need to coordinate their components generating an appropriate overall robot behavior that fulfills parallel goals of gathering scene information, achieving a task goal, communicate their internal status, and being always responsive to humans. This is especially relevant for complex scenarios in domestic settings.

Team of Bielefeld (ToBI) has been founded in 2009 and successfully participated in the RoboCup German Open from 2009-2013 as well as the RoboCup World Cup from 2009-2013. The robotic platform and software environment has been developed based on a long history of research in human-robot interaction [1–3]. The overall research goal is to provide a robot with capabilities that enable interactive teaching of skills and tasks through natural communication in previously unknown environments. The challenge is two-fold. On the one hand, we need to understand the communicative cues of humans and how they interpret robotic behavior [4]. On the other hand, we need to provide technology that is able to perceive the environment, detect and recognize humans, navigate in changing environments, localize and manipulate objects, initiate and understand a spoken dialog and analyse the different scenes to gain a better understanding of the surrounding.

In this year's competition, we extend the robot's capabilities for scene-analysis. Additionally to our furniture recognition system using Implicit Shape Model (ISM), we have added an Articulated Scene Model (ASM), which is able to segment functional parts of the scene in the robot's current view only from observation, while incorporating previously learned knowledge [5].

Another focus of the system is to provide an easy to use programming environment for experimentation in short development-evaluation cycles. We further observe a steep learning curve for new team members, which is especially important in the RoboCup@Home context. The developers of team ToBI change every year and are Bachelor or Master students, who are no experts in any specific detail of the robot's software components. Therefore, specific tasks and behaviors need to be easily modeled and flexibly coordinated. In concordance with common robotic terminology we provide a simple API that is used to model the overall system behavior. To achieve this we provide an abstract sensor- and actuator interface (BonSAI) [6] that encapsulates the sensors, skills and strategies of the system and provides a SCXML-based [7] coordination engine.

## 2 The ToBI Platform

The robot platform *ToBI* is based on the research platform *GuiaBot*™ by adept/mobilerobots[1] customized and equipped with sensors that allow analysis of the current situation. ToBI is a consequent advancement of the *BIRON* (**BI**elefeld **R**obot compani**ON**) platform, which is continuously developed since 2001 until now. It comprises two piggyback laptops to provide the computational power and to achieve a system running autonomously and in real-time for HRI.

The robot base is a PatrolBot™ which is 59cm in length, 48cm in width, weighs approx. 45 kilograms with batteries. It is maneuverable with 1.7 meters per second maximum translation and 300+ degrees rotation per second. The drive is a two-wheel differential drive with two passive rear casters for balance. Inside the base there is a 180 degree laser range finder with a scanning height of 30cm above the floor (SICK LMS, see Fig.1 bottom right). For controlling the base and solving navigational tasks, we rely on the ROS navigation stack[2].

In contrast to most other PatrolBot bases, ToBI does not use an additional internal computer. The piggyback laptops are Core i7 © (quadcore) processors with 8GB main memory and are running Ubuntu Linux. For person detection/recognition we use a 2MP CCD firewire camera (Point Grey Grashopper, see Fig.1). For object recognition we use a 13MP DSLR camera (Canon EOS 5D).

[1] www.mobilerobots.com

[2] `http://wiki.ros.org/navigation`

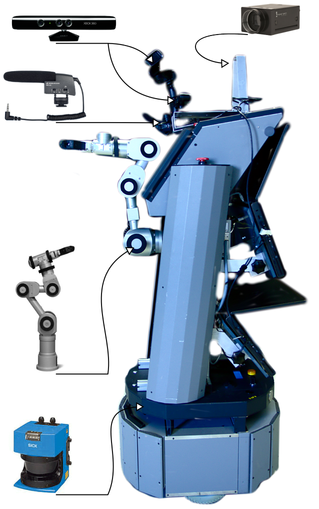

**Fig. 1.** ToBI with its components: camera, 3D sensors, microphone, KATANA arm and laser scanner.

For room classification, gesture recognition and 3D object recognition ToBI is equipped with an optical imaging system for real time 3D image data acquisition, one facing down (objects) and an additional one facing towards the user/environment. The corresponding computer vision components rely on implementations from Open Source libraries like OpenCV[3] and PCL[4].

Additionally the robot is equipped with the Katana IPR 5 degrees-of-freedom (DOF) arm (see Fig.1); a small and lightweight manipulator driven by 6 DC-Motors with integrated digital position encoders. The end-effector is a sensor-gripper with distance and touch sensors (6 inside, 4 outside) allowing to grasp and manipulate objects up to 400 grams throughout the arm's envelope of operation.

To improve the control of the arm, the inverse kinematics of the Katana Native Interface (KNI) was reimplemented using the Orocos [8] Kinematics and Dynamics Library (KDL)[5]. This allowed further exploitation of the limited workspace compared to the original implementation given by the vendor. This new implementation also enables the user to use primitive simulation of possible trajectories to avoid obstacles or alternative gripper orientations at grasp postures, which is important due to the kinematic constraints of the 5 DoF arm.

The on board microphone has a hyper-cardioid polar pattern and is mounted on top of the upper part of the robot. For speech recognition and synthesis we use the Open Source toolkits CMU Sphinx[6] and MARY TTS[7]. The upper part of the robot also houses a touch screen ($\approx 15in$) as well as the system speaker. The overall height is approximately 140cm.

## 3 Reusable Behavior Modeling

For modeling the robot behavior in a flexible manner ToBI uses the *BonSAI* framework. It is a domain-specific library that builds up on the concept of *sensors* and *actuators* that allow the linking of perception to action [9]. These are

[3] http://opencv.org/

[4] http://pointclouds.org/

[5] http://www.orocos.org/kdl

[6] http://cmusphinx.sourceforge.net/

[7] http://mary.dfki.de/

organized into robot *skills* that exploit certain *strategies* for an informed decision making.

We facilitate *BonSAI* in different scenarios: It is used for the robot BIRON which serves as a research platform for analyzing human-robot interaction [4] as well as for the RoboCup@Home team ToBI, where mostly unexperienced students need to be able to program complex system behavior of the robot in a short period of time. In both regards, the *BonSAI* framework has been improved such that system components are further decoupled from behavior programming and the degree of code re-use is increased.

To support the easy construction of more complex robot behavior we have improved the control level abstraction of the framework. *BonSAI* now supports modeling of the control-flow, as e.g. proposed by Boren [10], using State Chart XML. The coordination engine serves as a sequencer for the overall system by executing *BonSAI skills* to construct the desired robot behavior. This allows to separate the execution of the skills from the data structures they facilitate thus increasing the re-usability of the skills. The *BonSAI* framework has been released under an Open Source License and is available online[8].

## 4 Spatial Awareness

ToBI builds up different kinds of spatial representations of its environment using 2D and 3D sensors. This improves the robot's situation awareness and supports its searching abilities. In robotics, the role of egocentric representations of spatial information acquired through locomotion has been underestimated so far. In our *Articulated Scene Model* approach, we systematically deal with this aspect and present a method to generate a scene model of a system's current view incorporating past egocentric views by utilizing self-motion. Thereby, we exclusively rely on egocentric representations for perception tasks, while allocentric representations are used for navigation purposes and localization.

### 4.1 Articulated Scene Model

As a basis for the presented approach we use the *Articulated Scene Model* (ASM) introduced by Swadzba *et al.* [11]. This model enables an artificial system to categorizes the current vista space into three different layers: The *static background* layer which contains those structures of the scene that ultimately limit the view as static scene parts (e.g. walls, tables); second, the *movable objects* layer which contains those structures of the scene that may be moved, i.e. have a background farther perceived after moving (e.g. chairs, doors, small items); and third, the *dynamic objects* layer which contains the acting, continiously moving agents like humans or robots. An example is depicted in Fig. 2.

In the application for our robot ToBi, we focus on detection of completed changes which involves a comparison of currently visible structures with a representation in memory. Hence, our approach detects movable parts and adapts the

[8] `http://opensource.cit-ec.de/projects/bonsai`

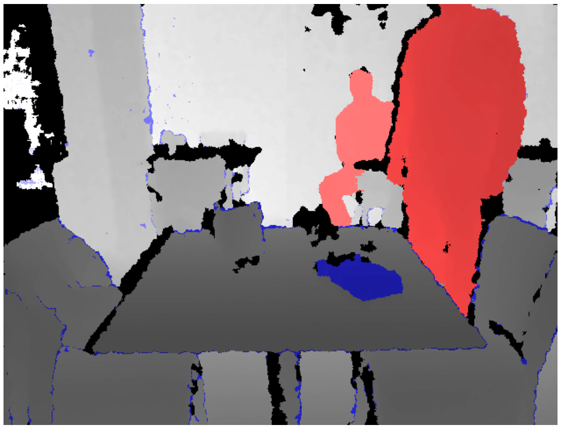

**Fig. 2.** Example for the Articulated Scene Model. Red: Dynamic. Blue: Movable

static background model of the scene simultaneously. The detection of dynamic object parts (like moving humans) is modeled in a separate layer and requires a tracking mechanism. For the implementation in the presented system, the body tracking algorithms in the NiTE Middleware[9] in combination with the OpenNI SDK[10] was used.

The original ASM algorithm works on depth images of a fixed scene. The background model is represented by a depth map that contains the greatest depth values for each pixel that were ever seen. Hence, the algorithm can assume that if a current depth reading for a certain pixel is smaller than the corresponding value in the background model, it must belong to a moveable object. This is because the observed obstacle is in a place which previously must have been clear because the background was visible. The tracking algorithm enables us to mark those movable parts of the scene model as dynamic, which are currently in motion.

### 4.2 Integrating multiple ego-centric models

In order to incorporate previously gathered information, the system needs to reliably fuse multiple scene models from different locations into a new model representing the current view. The method only transfers previously generated background models to the new scene, because movable parts can be calculated if the background is known. Because of the movable or even dynamic nature of the remaining parts it is likely that their location changed since the last observation. The distinction between movable and dynamic parts of the scene must therefore be done after merging the background models.

When initializing a scene model at a new location, all previously generated static background models (or a reasonable subset) are transformed into the current position of the camera using the memorized self-motion of the robot and

[9] `http://www.openni.org/files/nite/`, accessed 2014-02-06

[10] `http://www.openni.org/openni-sdk/`, accessed 2014-02-06

refinement through standard registration methods. For fusing the models with the current frame we developed a new merging algorithm which is similar to the basic articulated scene model algorithm, but utilizes the spatial relations of the merged models. As described before, a few special cases must be ruled out in order to ensure a correct model. If the transformed pixel does not meet the corresponding premises, it is ignored. The corresponding 3D point of the background pixel must have been in the **field of view** at the time when the incoming model was generated. Otherwise it is not safe to assume that the object was moved. It may just not have been visible. It must not have been **occluded** by any other point for the location of the camera of the incoming model, otherwise the object may again have been invisible because it was hidden. The candidate point must not have **neighboring points** in the incoming transformed model. This ensures stability to noise.

Figuratively speaking, the algorithm refines the currently perceived static background using evidences from other viewpoints and thereby fills areas that were not yet measured, e.g. because of shadows or reflecting surfaces. From the knowledge of the static parts of the scene at different times in the past, the algorithm can implicitly detect if an object was manipulated. Certain parts of the current scene will be marked as articulated if one of the merged models provides evidence that the corresponding object was not present at the time the model was built up or it was already known to be movable. The premises prevent that an object is falsely marked as movable because it was simply not visible from older views, but appeared in a subsequent view. So this algorithm allows the system to gain a much more informative articulated scene model without observing any change in the scene from the current viewpoint. Further details can be found in Ziegler *et al.* [5].

## 5 Implicit Shape Models

For robots acting in domestic environments the correct recognition of furniture objects is important for communicative and navigational tasks. During the RoboCup 2013 we have presented an approach based on Implicit Shape Models for recognizing pieces of furniture (see [12]). This approach is able to learn the spatial relationship of typical object regions from a set of artificial 3D models. For the detection of furniture objects in scenes captured with a Kinect camera, a 3-dimensional Hough-voting mechanism is used. One of the major drawbacks of this approach is that the training samples need to be truly scaled, which is often not the case when using public available data sets. Hence, a lot of work on model rescaling is necessary. Another drawback is the high computational effort during detection, being a cause of the 3-dimensional Hough-voting used. This effort even increases with the number of training samples, because a higher amount of training samples also leads to a higher amount of vote vectors assigned to the codewords. We tackled these two issues by changing the voting behaviour from a true 3-dimensional voting, including the vote direction and length, to a ray based voting only consisting of vote directions. This allows us to generate histograms of vote directions, instead of saving each individual vote vector. So the Implicit Shape Model now consists of a codebook and a specific vote direction histogram assigned to each codeword. During object detection, for each matched codeword vote rays are cast into the scene with vote weights relative to their corresponding histogram values. This allows us to use any scaled artificial model for learning, because only the directions from the detected regions to the object centroid are considered whereas the distances are omitted. Furthermore, the usage of a huge amount of training data does not influence the computational effort during detection any longer, because for each matched codeword the votes related to the histogram are cast, having a maximum in the number of histogram bins.

Furniture objects are finally detected in places showing a concentration of vote rays. Figure 3 shows example images of two scenes, their corresponding pointclouds and the detections received by our approach. For a detailed description of the implemented approach we refer the reader to Wittrowski *et al.* [13].

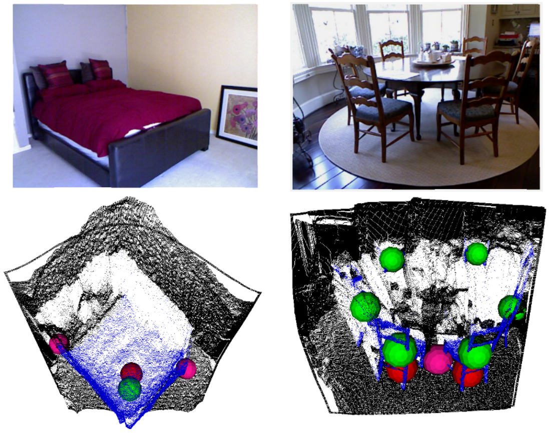

**Fig. 3.** The upper images show RGB-camera images. The bottom images show the point clouds and results. Green: ground truth; Red: true positives; Pink: false positives.

## 6 Conclusion

We have described the main features of the ToBI system for RoboCup 2014 including sophisticated approaches for the interpretation of 3D scenes. BonSAI represents a flexible rapid prototyping environment, providing capabilities of robotic systems by defining a set of essential skills for such systems. The RoboCup@HOME competitions in 2009 to 2013 served for as a continuous benchmark of the newly adapted platform and software framework. Especially BonSAI with its abstraction of the robot skills proved to be very effective for designing determined tasks, including more script-like tasks, e.g. 'Follow-Me' or 'Who-is-Who', as well as more flexible tasks including planning and dialog aspects, e.g.

'General-Purpose-Service-Robot' or 'Open-Challenge'. We are confident that the newly introduced features and capabilities will further improve the overall system performance.

## References

1. Wrede, B., Kleinehagenbrock, M., Fritsch, J.: Towards an integrated robotic system for interactive learning in a social context. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems - IROS 2006, Bejing (2006)
2. Hanheide, M., Sagerer, G.: Active memory-based interaction strategies for learning-enabling behaviors. In: International Symposium on Robot and Human Interactive Communication (RO-MAN), Munich (01/08/2008 2008)
3. Ziegler, L., Siepmann, F., Kortkamp, M., Wachsmuth, S.: Towards an informed search behavior for domestic robots. In: Domestic Service Robots in the Real World. (2010)
4. Lohse, M., Hanheide, M., Rohlfing, K., Sagerer, G.: Systemic Interaction Analysis (SInA) in HRI. In: Conference on Human-Robot Interaction (HRI), San Diego, CA, USA, IEEE (11/03/2009 2009)
5. Ziegler, L., Swadzba, A., Wachsmuth, S.: Integrating multiple viewpoints for articulated scene model aquisition. In Chen, M., Leibe, B., Neumann, B., eds.: Computer Vision Systems. Volume 7963 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 294–303
6. Siepmann, F., Ziegler, L., Kortkamp, M., Wachsmuth, S.: Deploying a modeling framework for reusable robot behavior to enable informed strategies for domestic service robots. Robotics and Autonomous Systems (2012)
7. Barnett, J., Akolkar, R., Auburn, R., Bodell, M., Burnett, D., Carter, J., McGlashan, S., Lager, T.: State chart xml (scxml): State machine notation for control abstraction. W3C Working Draft (2007)
8. Bruyninckx, H.: Open robot control software: the OROCOS project. In: Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164), IEEE (2001) 2523–2528
9. Siepmann, F., Wachsmuth, S.: A Modeling Framework for Reusable Social Behavior. In De Silva, R., Reidsma, D., eds.: Work in Progress Workshop Proceedings ICSR 2011, Amsterdam, Springer (2011) 93–96
10. Boren, J., Cousins, S.: The smach high-level executive. Robotics & Automation Magazine, IEEE **17**(4) (2010) 18–20
11. Swadzba, A., Beuter, N., Wachsmuth, S., Kummert, F.: Dynamic 3d scene analysis for acquiring articulated scene models. In: Int. Conf. on Robotics and Automation, Anchorage, AK, USA, IEEE, IEEE (2010)
12. Ziegler, L., Wittrowski, J., Schöpfer, M., Siepmann, F., Wachsmuth, S.: Tobi - team of bielefeld: The human-robot interaction system for robocup@home 2013 (2013)
13. Wittrowski, J., Ziegler, L., Swadzba, A.: 3d implicit shape models using ray based hough voting for furniture recognition. In: 3DV-Conference, 2013 International Conference on. (June 2013) 366–373