ToBI - Team of Bielefeld: The Human-Robot Interaction System for RoboCup@Home 2010

Sven Wachsmuth, Frederic Siepmann, Denis Schulze, Agnes Swadzba

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 the overall research goal. For robocup 2010, the team concentrates on mixed-initiative scenarios, person detection capabilities, and more sophisticated scene understanding methods.

1 Introduction

The Robocup@Home competition aims at bringing robotic platforms to use in regular home environments. There a robot needs to deal with unprepared domestic environments, autonomously perform in them and interact with laypersons. ToBI (Team of Bielefeld) has been founded in 2009 and successfully participated in the German Open 2009 (4th place, Hannover) and Robocup 2009 (8th place, Graz). 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 the 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. Thus, it is important to go beyond typical command-style interaction and to support mixed-initiative learning tasks. In the ToBI system this is managed by a sophisticated dialog model that enables flexible dialog structures [5].

In this year's competition, we extend the scene understanding and person detection capabilities of our robot. Most robotic systems build a 2D map of the environment by using laser scans and associate semantic labels to certain places that are known beforehand or are interactively teached like in the *walk-and-talk* task. However, there is no understanding of a table, desk, or sideboard where objects are typically placed on. In this year's Robocup@Home competition, we will make a first step towards this goal by integrating a 3D scene analysis component that is based on a Time-of-Flight depth sensor.

For person detection and tracking, we present a framework for the fusion of multiple cues and sensors. The main challenge is to deal with the correspondence problem on different time scales. Different cues have different processing times and provide partial results in different frequencies. At the same time it should be easy to add new cues. We present a flexible scheme that is based on a memory architecture.

A third focus of the system is to provide an easy to use programming environment for experimentation. Therefore, specific tasks and behaviors of the robot need to be fastly prototyped and iteratively changed during experimental trials. For this purpose, we provide an abstract sensor- and actuator interface (BonSAI) that encapsulates the sensors and components of the system.

2 Hardware

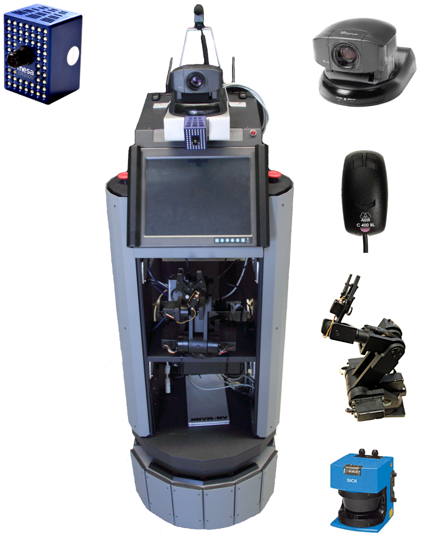


Fig. 1. The robot ToBI with it's components shown on the right. From top right: Pan-/Tilt-camera, interfacial microphone, Pioneer 5DOF arm and laser range finder.

The robot platform *ToBI* is based on the research platform *GuiaBot*$^{\text{TM}}$ by MobileRobots[1] customized and equipped with sensors that allow analysis of the current situation. ToBI is a consequent advancement of the former *BIRON* (**BI**lefeld **R**obot compani**ON**) platform, which has been under continuous development since eight years. 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$^{\text{TM}}$which is 59cm in length, 48cm in width, and 38cm in height, weighs approx. 45 kilograms with batteries and is maneuverable with 1.7 meters per second maximum translation and 300+ degrees rotation per second. It uses a two-wheel differential drive with passive casters for balance. Its foam-filled 19cm diameter wheels are at the center of rotation and it can carry a 12 kilogram payload. Inside the base there is a 180 degree laser range finder (SICK LMS, see fig.1). It can sense objects as far away as 50 meters with a ranging accuracy of 18 millimeters at a distance of up to 18m. The scanning height is at 30cm above the floor. The piggyback laptops are equipped with Intel Core2Duo©2GB main memory, running Linux. The camera is a 12x zoom pan-/tilt camera (SONY PTZ, see fig.1) that is able to scan an area of ±100 degree in front of the robot. For speaker localization two interfacial microphones are used (see Fig.1). For

[1] www.mobilerobots.com

the detection of planar surfaces and obstacles that are not in the laser range as well as for room perception, ToBI is equipped with an optical imaging system for real time 3D image data acquisition (Swissranger SR3000[2], see fig. 1). The camera features an integrated, modulated infrared light source that enables a time-of-flight based measurement, delivering a matrix of distance measurements independent from texture and lighting conditions.

Additionally the robot is equipped with a Pioneer 5 degrees-of-freedom (DOF) arm (see Fig.1); driven by six open-loop servo motors. The arms end-effector is a gripper allowing to grasp and manipulate objects as large as a can and as heavy as 150 grams throughout the arms envelope of operation. The upper part of the robot houses a touch screen ($\approx 15in$) and the system speaker. The overall height of the robot measures approx. 130cm.

3 System Architecture

For complex tasks as targeted by the RoboCup@Home challenge many different software components are in use that have to be orchestrated, coordinated, and integrated into one system. In order to provide the required flexibility, a cognitively motivated memory architecture serves as the foundation of our robotic system – both for the functional system architecture as also for the software architecture. This enables us to design the interactions within the system on a component based level (see sec. 3.1) as well as to model concrete system functionality on the application level (see sec. 3.2).

3.1 The Active Memory Service

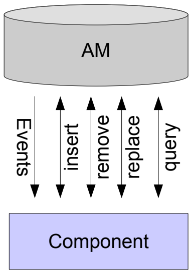


Fig. 2.

The *active memory* (AM) basically puts forward the concept of event-driven integration (EDI) on the basis of flexible event notification and XML-based representations as a document-oriented data model. In particular it comprises an “active memory service” (AMS) as a central integration broker for coordination and shared data management. Conceptually, all information generated and revised by components in the system is mediated through this active memory, where it can persistently be stored and retrieved from. The event-driven AM concept is directly supported by the Open-Source integration framework *XCF* [3][6]. It features adapters for the open-source computer vision framework *icewing* [4] employed by ToBI, and common robotic toolkits such as Player/Stage, MRPT[5], and others more.

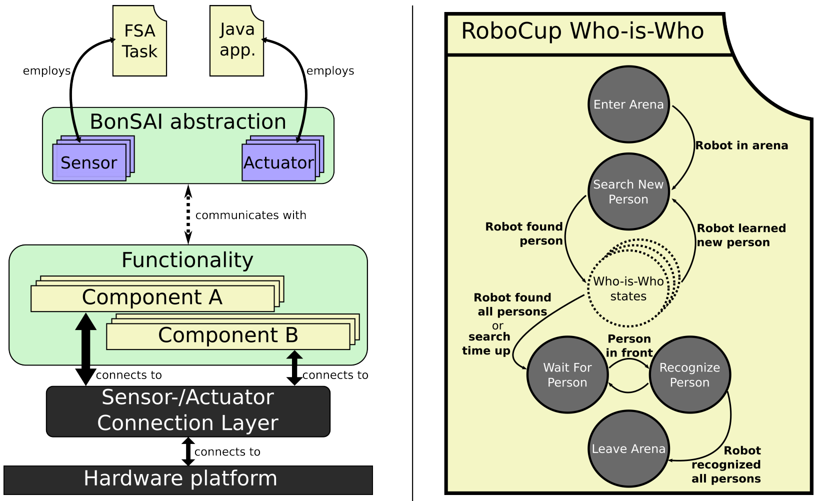
On top of this memory architecture, a functional API is defined that abstracts from specific components.

[2] Provided by the Swiss Center for Electronics and Microtechnology

[3] http://xcf.sf.net

[4] http://icewing.sf.net

[5] http://babel.isa.uma.es/mrpt



(a) BonSAI architecture concept

(b) Example for FSA-based application using BonSAI

Fig. 3. 3(a): The architecture as used for RoboCup@HOME. The Sensor/Actuator connection layer includes components used to connect to the hardware, functionality describes software components for sensing/acting, BonSAI abstraction is the collection of system functionality that are used by the FSA-based tasks (top left) or any other Java application (top right). 3(b): Example for an FSA-based application (Who-is-Who task RoboCup@HOME 2009).

3.2 Bonsai - sensor and actuator abstraction

The Biron sensor actuator interface (BonSAI) was developed to provide an abstract interface to the sensor information and actuators of the BIRON platform. The abstract Java API provides all hardware sensor information, e.g. laser data, encapsulated in a sensors class but additionally provides what can be called *cross-modal sensors* that may employ more than one hardware sensor or a combination of sensors and software components of the system. As depicted in fig 3(a) the functionality layer includes all software components of the system, e.g. face detection. The BonSAI abstraction layer includes all abstract sensors and actuators that are provided by the functional layer below. One example for a cross-modal sensor is the person sensor that returns information about persons detected by the system as described in section 4.1. Actuators in BonSAI can be distinguished in the same way: There are actuators that directly control the hardware, e.g. the Pan-/Tilt-/Zoom-Camera, but there are also cross-modal actuators such as the *NavigationActuator* that employs different components of the system to get the robot to a certain location. The BonSAI abstraction layer enables us to define a system functionality that can be used by applications or components which make use of all the components of the BIRON platform. Apart from a rapid prototyping environment for the robot and its functions this provides the basis for the participation in the RoboCup@HOME. Each of the applications for the RoboCup tasks implements a finite state automata (FSA) where all sensor information can trigger state transitions. For timing constrains all FSA-based tasks have an additional timer that can trigger state transitions

[5] http://playerstage.sourceforge.net

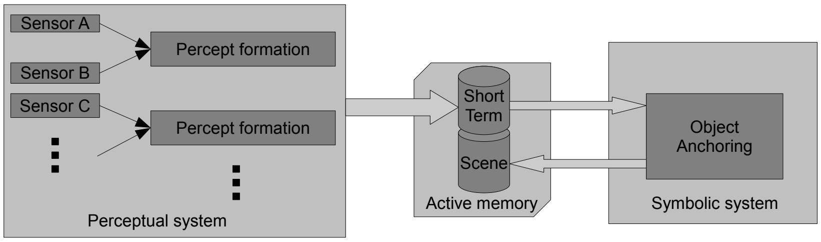


Fig. 4. Schematic memory-based information flow. Percepts are generated by different modules in the perceptual system and are inserted into the short term memory. The anchoring system receives the new percepts and inserts a document for anchored persons into the scene memory.

as well. In fig. 3(b) the *Who-is-Who* task from the 2009 RoboCup@HOME is described in more detail. 3(a) also shows is that there are no direct connections to the system hardware. This is transparent in BonSAI via the functionality layer and it leads to a high re-usability of the tasks or applications using BonSAI. Some first experiments indicate that it is possible to exchange the hardware and use the same FSA-based tasks from the RoboCup on a different system without changing anything in the task. If a task uses ie.g. the person sensor, we simply provide the information necessary for this sensor, independent of the hardware or component providing such information.

4 Software Components

Although the software architecture is organized in terms of components or *memory processes*, the functional skills of the system emerge from the interaction of multiple components that exchange information via the active memory. In the following, we will describe some of skills that are accessable as sensors or actuators in the BonSAI-API.

The localization and navigation is based on an open source component[6] which uses the bubble band approach with integrated obstacle avoidance. During the guided tour SLAM is used in combination with data from the laser range finder and wheel odometry to create a map. The implemented interactive location learning [7] integrates symbolic information like room or object labels into human-augmented maps to facilitate autonomous navigation afterwards.

The dialog system is based on the grounding of multi-modal interaction units [5]. It gets information about the current system state from the memory and combines the information with the user utterances processed by a speaker-independent speech recognizer and speech understanding system.

4.1 Person Tracking

For the person tracking a memory-based multi modal anchoring system is used. The main purpose of any anchoring system can be described as ”the process of creating and maintaining the correspondence between symbols and percepts that refer to the same physical object” [8]. In principle the anchoring system consists

[6] http://libsunflower.sf.net

of two different layers that are interconnected using the active memory system. The first layer (see fig. 4 left) is the perceptual system. This system includes a list of percepts where each percept is a structured collection of measured attributes. The second layer is the symbolic system (fig. 4 right). This layer includes abstract descriptions of objects in the physical world, so called symbols. The link between a percept and a symbol is a data-structure called anchor. It includes the percept, the symbol and a signature that provides an estimation for the objects' observable properties. In a dynamic environment an anchor has to be updated every time new percepts are generated so that an anchor is a dynamic structure indexed over time.

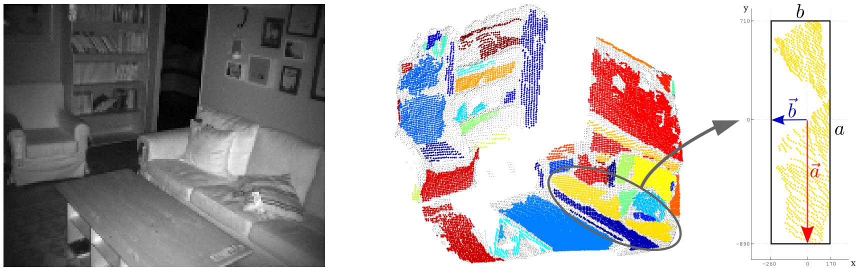
The classical approach describes the special case of connecting one percept to one symbol. Nowadays robots are often equipped with different sensors that can measure different properties of an object. For example a laser range finder can detect the distance of a person to the robot more precisely than a face detector. To include these different sensors, the classical approach is extended into a multi-modal anchoring approach, where different anchors can link different percepts to the symbols. These anchors are first combined into a composite anchor that constitutes the link to the symbol in the symbolic system [9].

The memory-based scheme presented in Fig. 4 leads to a decoupling of different modalities that each have access to the different percepts generated by the perceptual system. This enables the possibility to change fusion strategies without the need to change the tracking heuristics.

For the person tracking, the robot uses the information from the laser range finder to detect the distance and angle of person legs with regard to the robot. Additionally, a second estimation of a persons distance and angle is calculated using a face detector. When both, a face percept and a leg percept, are linked together to a symbol person the two different values are fused together to provide a better estimation of the position of the person relative to the robots coordinate system. Due to the fact that the different sensors work at differents frequencies, only percepts that contain a timestamp not too far away from the current system time are taken into account.

4.2 3D scene analysis

The methods mainly used to acquire 3D information can be divided in passive (e.g. stereo vision) and active methods (e.g. laser range scanners or Time-of-Flight (ToF) sensors). However, stereo vision depends on the environmental conditions, this means the appearance of the scene strongly influences the quality of the point cloud generation. Active sensors overcome this restriction by generating and sending a signal on their own and measuring the reflected signal. 3D Time-of-Flight (ToF) Sensors [10] combine the advantage of active sensors and camera based approaches as they provide a 2D intensity image and exact distance values in real-time. Compared to stereo rigs the 3D ToF sensors can deal much better with prominent parts of rooms like walls, floors, and ceilings even if they are not textured.



(a) amplitude image

(b) 3D point cloud and example plane

Fig. 5. This figure shows an exemplary output of the Swissranger camera (frame 162 of `liv.5`) – (a): the amplitude image, (b): the 3D point cloud with points belonging to the same planar surfaces highlighted by the same color and a plane transformed so that the vectors indicating the two largest variance directions in the data are parallel to the coordinate axis.

As man-made environments mostly consist of flat surfaces a simplified representation of the scene is generated by extracting bounded planar surfaces from the 3D point cloud delivered by the Swissranger camera (5).

First, noisy data from the camera is smoothed by applying median and edge filters to the depth map of each frame. The decomposition of the 3D point cloud into connected planar regions is done via region growing as proposed in [11] based on the conormality measurement defined by [12]. Iteratively, points are selected randomly as seed of a region and extended with points of its 8-neighborhood if the points are valid and conormal. The resulting regions are refined by some runs of the RANSAC algorithm. Figure 5(b) displays for a frame the resulting set of planes.

The collection of scene planes can be used for different purposes. On the one hand, planes at specific heights may define potential tops of tables and sideboards. Those at lower heights might be a seating plane of chairs or sofas. This can be used to navigate around obstacles that cannot be seen by the laser or constrain the search for objects requested. On the other hand, the statistical distribution of planes can be used in order to categorize different parts of a room like dining area or kitchen. Therefore, a histogram vector is computed considering the shape, size, and orientation of plane pairs and is classified for the pre-learned room categories [11].

5 Conclusion

We have described the main features of the ToBI system for Robocup 2010 including sophisitcated approaches for person detection and 3D scene analysis. Bonsai represents a flexible rapid prototyping environment, providing capabilities of robotic systems by defining a set of essential functions for such systems.

The RoboCup@HOME competition in 2009 served for as an initial benchmark of the newly adapted platform. The Team of Bielefeld (ToBI) finished 8th place, starting with the new hardware and no experience in competitions like RoboCup. The determined tasks had to be designed from scratch because there where no such demands for our platform prior to the RoboCup competition. BonSAI with its abstraction of the system functionality proved to be very effective for designing determined tasks, e.g. the Who-is-Who task where the robot has to autonomously find three persons in the arena and re-identify them at the entrance door of the arena in a given time. This scenario is well defined for a script-like component as the number of people in the scene is known in advance and also what actions the robot should take. Additionally the runtime of the task can be used as ultimate trigger for the robots behavior. In contrast open challenges have no determined set of goals, the robot can show basically anything it is capable of. This leads to an open scenario where all capabilities where shown in an interactive manner which means there needs to be a interaction with the user.

References

1. Haasch, A., Hohenner, S., Hwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G.A., Fritsch, J., Wrede, B., Sagerer, G.: Biron – the bielefeld robot companion. In: Proc. Int. Workshop on Advances in Service Robotics. (2004) 27–32
2. 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)
3. 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)
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, IEEE (11/03/2009 2009)
5. Li, S., Wrede, B., Sagerer, G.: A computational model of multi-modal grounding. In: Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006, ACL Press, ACL Press (2006) 153–160
6. Fritsch, J., Wrede, S.: An integration framework for developing interactive robots. In Brugali, D., ed.: Springer Tracts in Advanced Robotics. Volume 30. Springer, Berlin (2007) 291–305
7. Peltason, J., Siepmann, F.H., Thorsten P. Spexard, B.W., Hanheide, M., Topp, E.A.: Mixed-initiative in human augmented mapping. In: Processings Int. Conference on Robotics and Automation. (2009) to be published.
8. Coradeschi, S., Saffiotti, A.: Perceptual anchoring of symbols for action. In: Proc. of the 17th IJCAI Conference, Seattle, Washington (2001) 407–412
9. Fritsch, J., Kleinehagenbrock, M., Lang, S., Pltz, T., Fink, G.A., Sagerer, G.: Multi-modal anchoring for human-robot-interaction. Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems **43**(2–3) (2003) 133–147
10. Weingarten, J., Gruener, G., Siegwart, R.: A state-of-the-art 3D sensor for robot navigation. In: Proceedings of the International Conference on Intelligent Robots and Systems. Volume 3. (2004) 2155–2160
11. Swadzba, A., Wachsmuth, S.: Categorizing perceptions of indoor rooms using 3d features. In: Lecture Notes in Computer Science: Structural, Syntactic, and Statistical Pattern Recognition. Volume 5342. (2008) 744–754
12. Stamos, I., Allen, P.K.: Geometry and texture recovery of scenes of large scale. Computer Vision and Image Understanding **88**(2) (2002) 94–118