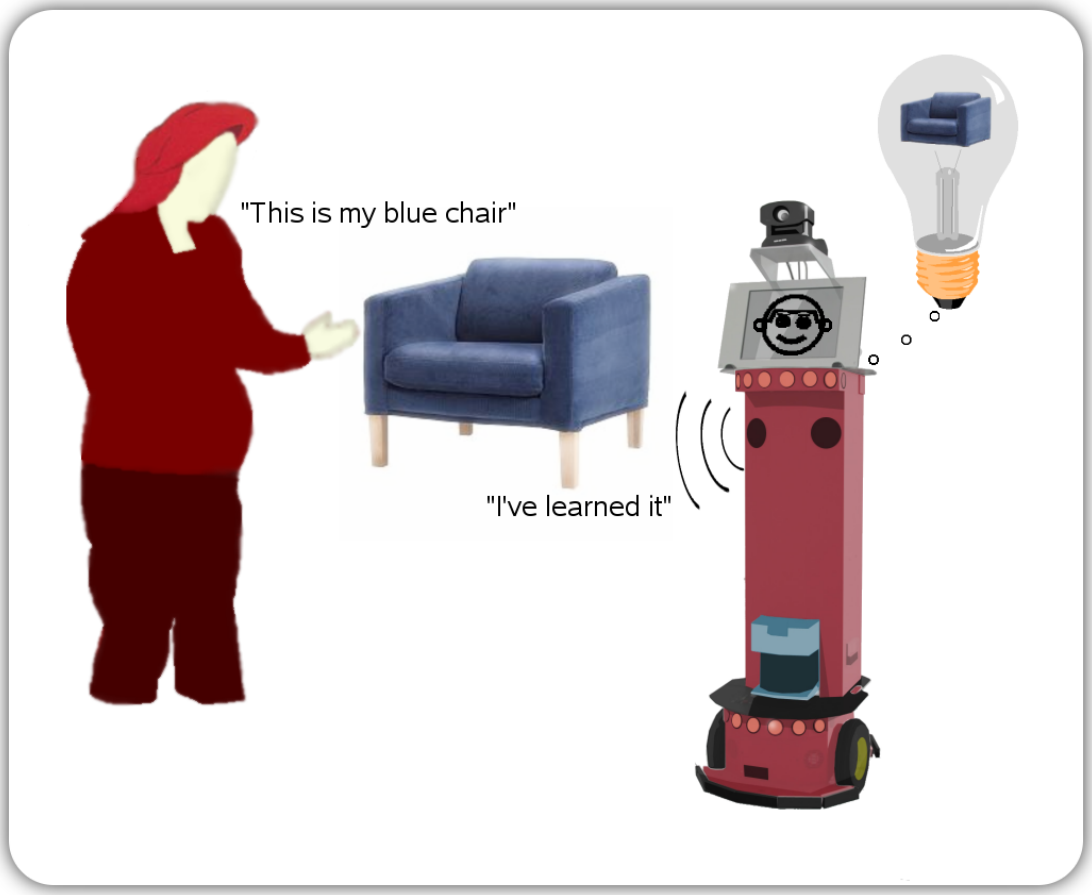
Attention-controlled Acquisition of a Qualitative Scene Model for Mobile Robots

Axel Haasch



Universität Bielefeld
Angewandte Informatik

Attention-controlled Acquisition of a Qualitative Scene Model for Mobile Robots

Der Technischen Fakultät der
Universität Bielefeld

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Axel Haasch

26. Januar 2007

Dipl.-Inform. Axel Haasch
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
E-Mail: ahaasch@TechFak.Uni-Bielefeld.DE

Abdruck der genehmigten Dissertation zur Erlangung des
akademischen Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.).
Der Technischen Fakultät der Universität Bielefeld
am 26.01.2007 vorgelegt von Axel Haasch,
am 22.05.2007 verteidigt und genehmigt.

Gutachter:
Prof. Dr. Helge Ritter, Universität Bielefeld
Dr. Jannik Fritsch, Honda Research Institute Europe, Offenbach

Prüfungsausschuss:
Prof. Dr. Helge Ritter, Universität Bielefeld
Prof. Dr. Franz Kummert, Universität Bielefeld
Dr. Marc Erich Latoschik, Universität Bielefeld
Dr. Jannik Fritsch, Honda Research Institute Europe, Offenbach

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit mit dem Titel

Attention-controlled Acquisition of a

Qualitative Scene Model for Mobile Robots

selbstständig verfasst, sinngemäß sowie wörtliche Zitate kenntlich gemacht und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Bielefeld, 26. Januar 2007

Axel Haasch

Abstract

Robots that are used to support humans in dangerous environments, e.g., in manufacture facilities are established for decades. Now, a new generation of service robots is focus of current research and about to be introduced. These intelligent service robots are intended to support humans in everyday life. To achieve a most comfortable human-robot interaction with non-expert users it is, thus, imperative for the acceptance of such robots to provide interaction interfaces that we humans are accustomed to in comparison to human-human communication. Consequently, intuitive modalities like gestures or spontaneous speech are needed to teach the robot previously unknown objects and locations. Then, the robot can be entrusted with tasks like, fetch-and-carry orders even without an extensive training of the user. In this context, this dissertation introduces the multimodal Object Attention System which offers a flexible integration of common interaction modalities in combination with state-of-the-art image and speech processing techniques from other research projects. To prove the feasibility of the approach the presented Object Attention System has successfully been integrated in different robotic hardware. In particular, the mobile robot BIRON and the anthropomorphic robot BARTHOC of the Applied Computer Science Group at Bielefeld University. Concluding, the aim of this work to acquire a qualitative Scene Model by a modular component offering object attention mechanisms has been successfully achieved as demonstrated on numerous occasions like reviews for the EU-integrated Project COGNIRON or demos.

Contents

1. Introduction

"Please clean up the table and bring the dishes into the kitchen."

The sentence afore might in the future become a typical utterance of an elderly person directed to a robotic companion. This addresses the development that even today more and more elderly people decide to live alone and, hence, in some circumstances require external help. The situation illustrates a growing problem due to the foreseeable shift of the age pyramid (Figure 1.1) [fE04].

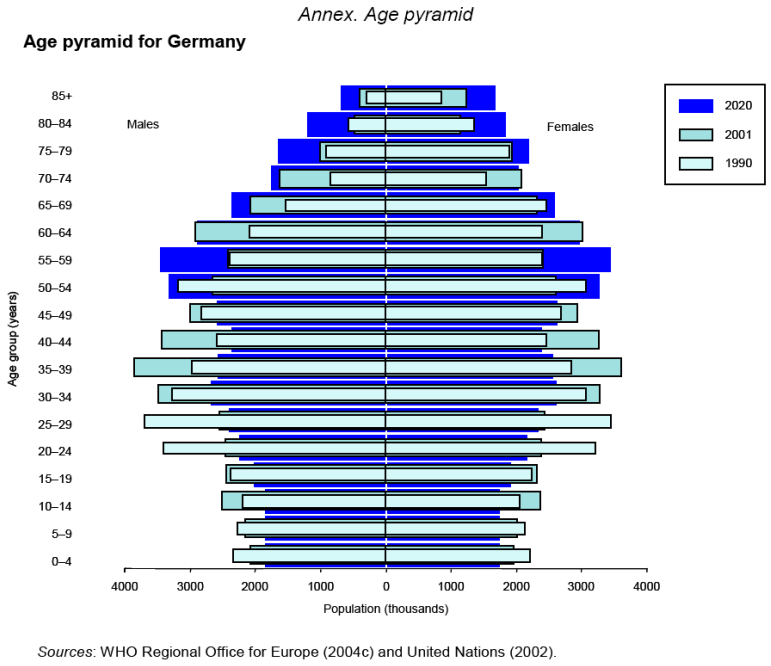


Figure 1.1: Age pyramid for Germany.
©2004 WHO. The image has been taken from [fE04].

Significant progress in medical issues and a regressive birth rate are the most important reasons for this development in our modern society. In the end this will

lead to the problem that more and more people have to be cared for at home as the capacity of retirement homes will exceed. Therefore, the society is on search for a solution that enables home care as, in this way, the people can stay in their familiar living environment. Consequently, a lot of research is done in the field of robotics in order to match these requirements of a home care involving a robot as caregiver. Hence, the development of so-called *Personal Robots* which represent companions that are able to learn from humans is one of the major goals of this research. In particular, the establishing of a natural *Human-Robot Interaction (HRI)* that offers a learning aspect in terms of adaptations in the robot's behavior on the needs of the interaction partner is the most difficult issue within the field of robotics. This includes not only the interfaces that are used during such an interaction but also the robot's capability to appropriately answer the queries stated by the user. For instance, social aspects like upholding a certain distance to the user or the character of the robot (introvert vs. extrovert) need to be considered as well. For such *Robot Companions* [DWK+05] or *Robot Assistants* [LPD+01][1] numerous applications are conceivable. Not only health care but also surveillance and entertainment tasks or even support for pregnant women, cf. cover illustration.

In combination with the development of a fully automated domestic *Personal Robot* many problems in vision and speech processing have to be solved. To face these challenges the *Robot Home Tour Scenario* was introduced by the European *Cognitive Robot Companion project (COGNIRON)* [COG06] (Figure 1.2). In particular, the scenario concentrates on the learning process of a robot after the user has bought it in a store and took it home. As soon as the user switches the robot on for the very first time, it automatically initializes and is, thus, ready to acquire knowledge about its new operating area that contains, for instance, various objects. Because everything is unknown to the robot in this state, the user has to teach the robot all objects and locations that might become important for subsequent autonomous interactions.

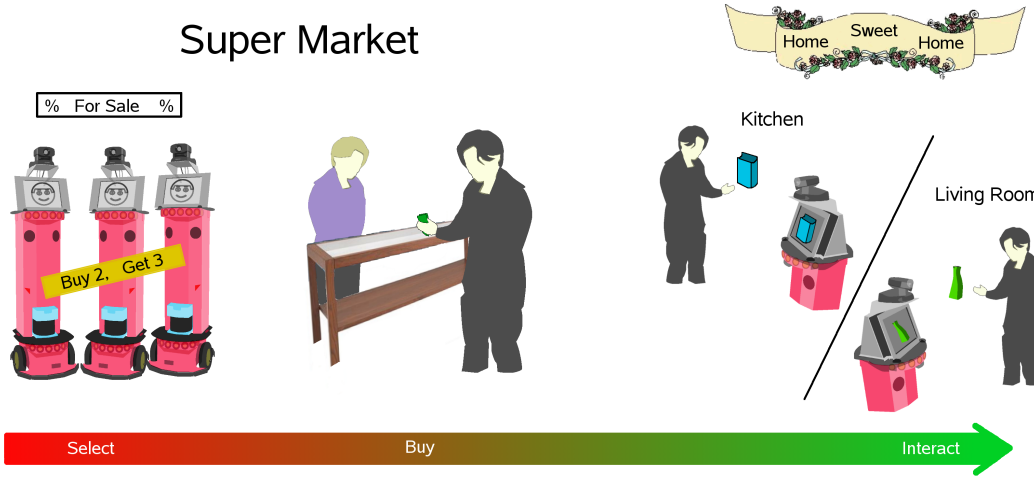


Figure 1.2: The *Robot Home Tour Scenario*.

In addition to the learning of unknown object instances it is useful that the robot is able to learn the geometry and the topology of the environment as well. Thus,

[1] cf. MORPHA-project [MOR99] and Desire-project [fPuA05]

autonomous interactions with its environment become easier. Consequently, an analysis of spatio-temporal relations between different objects can help to match the peculiarities of a dynamic environment, like a private home with, e.g., objects that can be carried to various locations or rooms. Furthermore, the robot has to be able to gain knowledge about its user, e.g., the name, for the purpose of a comfortable communication.

While dealing with the large amount of technical problems resulting from these requirements, however, the important issue whether the society is already willing to accept a *Personal Robot* as a companion needs to be considered as well. Within this context, Arras and Cerqui conducted an opinion survey with 2000 participants [AC05]. Although, the study was restricted to one European country, the answers clearly show that people still exhibit a lot skepticism regarding the acceptance of robots in their life. Thus, significant clarification of the benefits of a robotic companion is needed. Additionally, the intuitive handling of robots has to be improved for interactions with non-expert users. Especially the ease of use is an extremely complex task and consists of a lot of requirements that have to be fulfilled. But as shown in the past the learning aspect is the most challenging one among all tasks, because it is a highly adaptive process. It is impossible to foresee all situations due to the unstructured, cluttered, and flexible environmental conditions. Furthermore, elderly people are usually not as familiar with new robotic techniques as young people. This makes a simple, flexible, and adaptive interface absolutely essential for a successful Human-Robot Interaction and motivates us to search for new integrated solutions.

Motivation

The learning of unknown objects during a Human-Robot Interaction demands for attentional mechanisms that enable the robot to steer its attentional focus on individual object locations.

This especially applies for assistant-like personal robotic companions that are intended to be used for tasks requiring interactions with their user and their environment, like, e.g., fetch-and-carry tasks ("bring me tea"), cleaning tasks ("clean up the table"), or search tasks ("find my keys"). However, the robots can only solve this task if they are able to perceive and analyze objects and locations in their vicinity. Decades of research have shown that robots with a perception only based on a single modality are often not suitable for these tasks. Hence, the logical conclusion is to use different modalities in parallel, e.g., vision and sound, in order to overcome the limitations of former approaches in the field of object perception for mobile robots. But, even such integrated solutions are not yet able to satisfy the needs of non-expert users in natural environments as they are usually developed from a technical perspective only and, thus, are not able to communicate in terms that humans are accustomed to. That is the reason why recent cognitive robotic research considers not only the object itself, but also the user as the most valuable source of information.

In doing so, a selection of relevant input that can be influenced by, e.g., the utterances or the gestures of the currently interacting user becomes useful. This selection is often realized by using an attentional mechanism [Lan05]. In particular, this means that the robot needs to have an awareness model for objects

that enables the robot to focus its attention on the same *Region-Of-Interest* that the user refers to. It is easy to understand that this process has to be as comfortable and intuitive for the human as possible. Hence, the awareness model needs to support different modalities that humans are accustomed to. Usually this involves at least the use of speech and deictic gestures. Due to their close relationship between each other, especially these two modalities need to be analyzed and aligned in a spatio-temporal sense. Thus, object references specified by the user can be resolved which enables the robot to determine the current Region-Of-Interest. This region is, consequently, used to acquire detailed visual and auditory information about an object. As an outcome of such a successful interaction, the gained knowledge can then be used for the acquisition of a qualitative *Scene Model* that stores all relevant information. This not only enables the robot to interact with its environment on its own, but also ensures the evaluation of the user's actions through continuous feedback and provides the user and the robot with a common ground. Hence, ambiguous situations between the user and the robot can be avoided by, e.g., displaying an image about the selected object, and yielding a more comfortable interaction. Concluding, the user always needs to be informed about the internal state of the robot during the analysis phase in order to be able to recognize and, eventually, to influence the robot's behavior.

This internal state has to be communicated as best as possible. The probably best choice, therefore, seems be an adaption of the robot's behavior and attention to human standards, respectively. The most promising attention models to face these challenges are, consequently, presented in the following as far as they are relevant in this work.

Survey of Attention Models for Cognitive Robots

The notion *Attention* is difficult to interpret. Mostly, this is due to its abstractness and the unmanageable diversity of possible definitions. In this thesis definitions of Attention given in [KH04] and [Nag04] are used. Accordingly, Attention is considered as a process connected to an agent which concentrates on some features of the environment while other features are excluded.

Here, the focus lies on modalities that are usually used for a Human-Robot Interaction. To uphold the close relation between the user and the robot, and because humans are excellent in paying attention on distinctive features, all following theories of attentional mechanisms are biologically inspired. Therefore, they are regarded in the context of so-called *Social Attention* since at least two individuals are involved in a Human-Robot Interaction. Social Attention in terms of visual perception, as it is described in [Eme00] can be distinguished into fives classes: *Mutual Gaze*, *Gaze Following*, *Joint Attention*, *Shared Attention*, and *Theory of Mind*. As Mutual Gaze and Gaze Following are rather considered as preprocessing for the establishment of an attention-based *Human-Robot Communication*, it is useful to concentrate on the latter classes which are more feasible. In this work these different classes of attention between two interaction partners are understood as follows:

- **Joint Attention** between the user and the robot defines a joint intentional relation to the world pursuing a plan of action that either the user or the robot chooses in order to realize a particular goal.

- **Shared Attention** regarded as extension of Joint Attention additionally describes that the human and the robot are aware of each other and know that the partner focuses its attention on the same feature.
- **Theory of Mind** describes the ability of the robot to attribute the behavior of its user to the user's mind, like it is often done between humans as well [BC95]. For instance, when the user grasps an apple, the robot could infer that the user is hungry and probably wants to eat the apple.

Starting from these definitions, the following section points out the contribution of this thesis for the realization of an attentive system for objects supporting intuitive interactions between a human user and a robot.

Contribution

The contribution of this thesis is the development of an innovative approach for the integration of different modalities in an attention-controlled framework for object learning and recognition tasks that emerge from a Human-Robot Interaction. The proposed framework is called *Object Attention System (OAS)*, in the following used as proper name that accurately establishes a spatio-temporal relation between visual and auditory object features, symbolic speech input, face recognition, and gesture information. Thus, a so far unique amount of acquired object information becomes available which extends the imaginable applications for the Object Attention System in numerous ways. For instance, intuitive Human-Robot Interactions with object references become possible due to interfaces that humans are accustomed to (natural speech, pointing gestures, ...). Additionally, the robot can interact with its environment autonomously, like reacting on the sound of an alarm clock even it can not be seen by the robot while afterwards the robot could inform its user, e.g., by a spoken notification.

In order to face these challenges, the Object Attention System integrates many recent techniques that are summarized next. Resulting from interactions, the object data includes graph-based visual appearances [TE05] as well as *SIFT features* [Low04] to support a proper object recognition of objects afterwards. From an auditory perspective, modern compression algorithms, like *Ogg Vorbis* [Fou06] are used to achieve an adequate sound representation of objects as well. Especially the object representation, no matter whether it concerns sound, text, or vision, consistently uses broadly accepted open and non-proprietary formats only, like *OpenCV* [int06] or *XML* [Wor06a].

Additionally, the symbolic description given by the user is summarized in an adaptive XML-based structure which, for instance, includes 3D-object positions, object properties, and entries pointing to the location of object sounds and object views. In conclusion, the gathered information is used to acquire a qualitative *Scene Model* representing the environment. To prove the feasibility of the proposed approach, it has been applied on two different demonstration platforms, the mobile robot *BIRON* [HHH+04] and the anthropomorphic robot *BARTHOC* [HSF+05]. Therefore, the Object Attention System has been integrated in the robotic system infrastructure *SIRCLE* [FKH+05]. To sum up, the proposed Object Attention System meets the demands for a natural, comfortable, and intuitive Human-Robot Interaction.

Outline

The complexity of the contributed Object Attention System is explained best by Figure 1.3 on page 7. The cyan framed numbers relate to the relative chapter of this thesis. Thus, this thesis is structured as follows.

Chapter 2 presents related work in the context of Object Attention used in robots. In particular, the chapter gives a more detailed introduction in the concepts of Joint Attention, Shared Attention, and the Theory Of Mind. Then, a reflection of unimodal versus multimodal attention-based systems follows. At the end of chapter 2, an overview on robots using attentional mechanisms is given.

The subsequent chapter 3 first discusses possible modalities that are useful for the development of the multimodal Object Attention System, followed by a view on the modalities actually used. Thus, a distinction between user-related and object-related perception, as well as supported interfaces (e.g., symbolic speech, gesture data) are described.

Next, after the modalities have been introduced, chapter 4 focuses on the development of the integrated *Object Attention System*. Therefore, a short section about related work with regard to learning approaches for objects and the robot hardware used is provided. This is followed by a detailed presentation of the integration of intra- and intermodule-relevant communication, while time-dependent restrictions are considered as well. After that, the acquisition of an appropriate Region-Of-Interest together with the realized object representation of the user's actions and visual as well as acoustic object appearance is described. The chapter closes with the presentation of the implemented processing strategies for the handling of unknown and known objects.

In order to complete the integrational aspects of the proposed Object Attention System, chapter 5 gives an insight into the robotic system environment where it has been integrated. Therefore, the chapter deals with the knowledge representation in terms of a Scene Model which is realized by an *Active Memory* concept, and a method to gain enhanced visual object perception by so-called *Multi-Mosaic images*. In the second half of that chapter, the actual system environment is described, ranging from the architectural framework to the communication system applied. So far, all aspects for the development of the Object Attention System have been presented. Thus, the developed system needs to prove its technical quality.

The quality analysis of such a system, hence, is the topic of the following chapter 6. Here, a qualitative and quantitative evaluation of the proposed object reference resolving approach is given. Therefore, an extensive evaluation setup including user studies is presented. The chapter ends with a discussion of the acquired evaluation results.

Finally, a summary of the work including an outlook for possible extensions of the proposed Object Attention System is given in chapter 7.

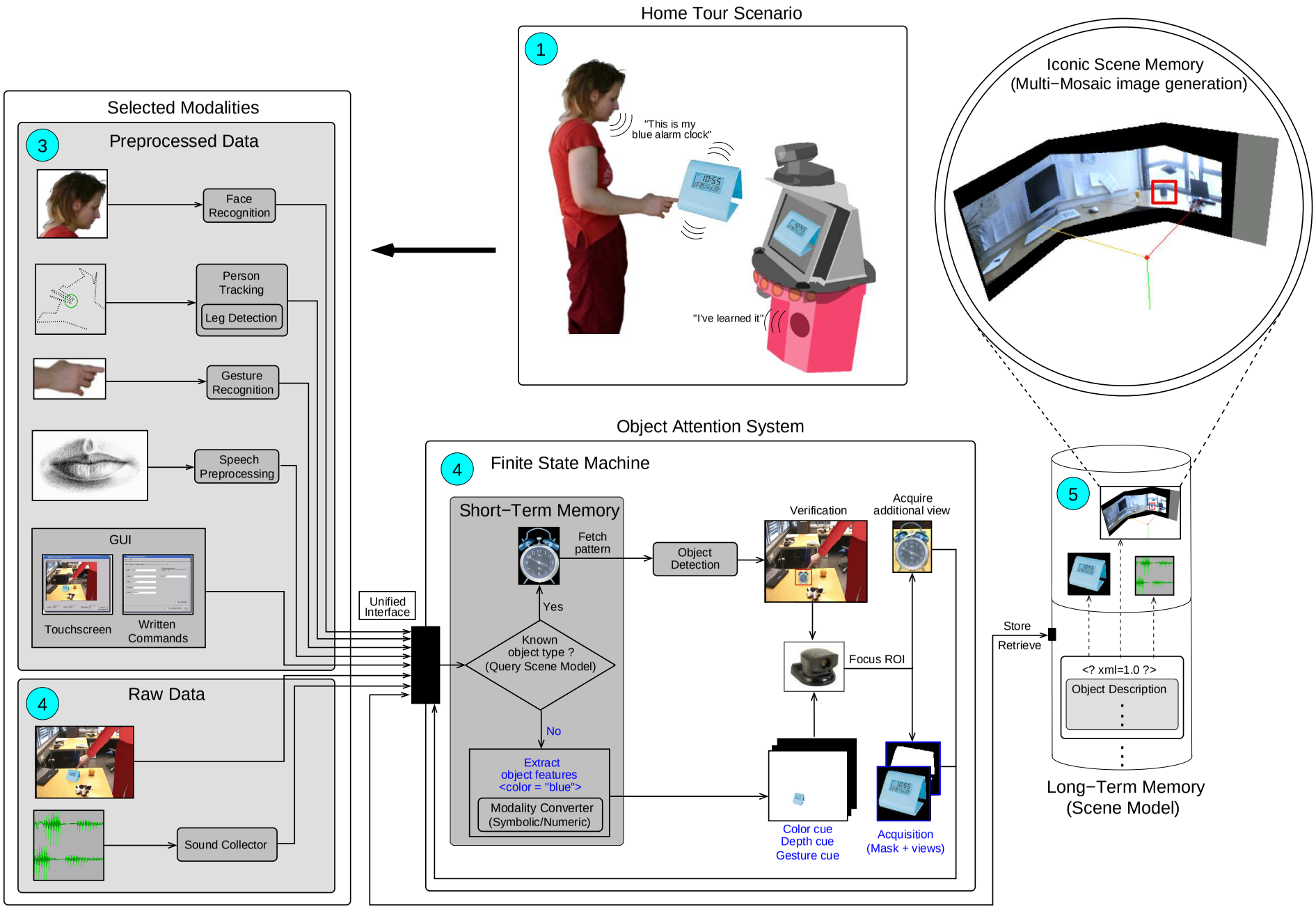


Figure 1.3: Component diagram for visualization of the thesis outline.
cyan-framed numbers denote the chapters of this thesis (chapter 2 is not listed as it contains related work).

2. Object Attention and Learning in Mindful Robots

Attention as intention to concentrate on a distinctive feature is an essential part of an interaction or communication. Hence, sophisticated Personal Robots that are intended for interactions with a human, need to have the ability to direct their attention on objects the human user refers to. Additionally, as this kind of robots is meant to be used as companion-like robots, they have to support human-like behavior as well, e.g., simultaneous looking on objects for feedback reasons or verbal feedback after an object has been focused.

This chapter begins by describing an attentive robot equipped with a basic attention model in order to illustrate the principles of directed *Object Awareness*. Subsequently, more advanced models are presented which follow the same principles as proposed in the first example. These more sophisticated approaches, namely Joint Attention, Shared Attention, and Theory Of Mind are then discussed within unimodal versus multimodal robotic architectures providing Object Attention. At the end of this chapter a selection of robots that use mechanisms which allow them to interact with objects are presented.

Kopp and Gärdenfors propose in [KG01] that "...attention is a minimal criterion of intentionality in robots...". This is a clear statement, because only a robot that is able to perceive its environment is able to act in it. To prove their point they have developed the reactive grasp robot R1 in order to describe behavior that seems intentional. In particular, R1 uses a set of S-R (stimulus-response) rules. For grasping a single object they have observed that R1's action seems to be intentional since the robot is able to appropriately react even on moving objects. However, using S-R rules limits the interactions to single objects. If more than one object is present, the "attention" of R1 randomly shifts between the different objects.

Considering their experience with R1 they submit that a visual robot needs to be able to

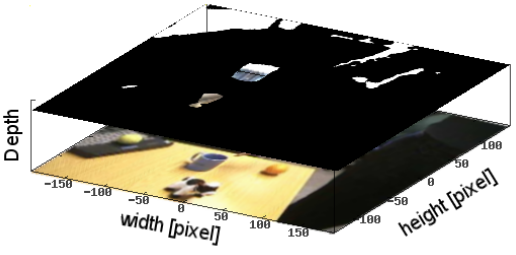
1. *identify* relevant objects in the scene.

2. *select* one of the identified objects.
3. *direct* its sensors towards the selected object.
4. *maintain* its focus on the selected object.

These prerequisites lead to their proposal of a basic architecture for an attentional robot consisting of a reactive system, a value component to represent the motivational attraction level for objects, a selection system and an attentive system. To sum up, this is the basis for an attentional mechanism used for object learning and recognition in a robot. However, for an intuitive Human-Robot Interaction not only salient object features need to be analyzed but also the user's instructions. Here, humans are mostly accustomed to a combination of speech and gestures because gestures usually provide a faster and more precise possibility to refer to objects. For instance, it is often very difficult to describe the exact position of an object verbally only. Thus, adding visual gesture information can help to improve the convenience of an interaction. But, this results in the need for a combination with the visual appearance of the referenced object. Most commonly the combination of different visual input is realized with some kind of *Attention Maps* [SRHR04] (similar to *Conspicuity Maps* [IKN98], *Feature Maps* [BS99] or *Saliency Maps* [VCSS01]) while their basic idea is exemplary depicted in Figure 2.1(b).

(a) Original input image captured by the object camera.

(b) A color mask for 'blue'. White areas are transparent.



(c) Overlay schematics for highlighting blue color (isometric view).

(d) Overlay schematics for highlighting blue color (top view).

Figure 2.1: Illustration of the usage of possible cues for Attention Maps.

Those Attention Maps highlight a specific feature and, ideally, provide only data that is relevant for the evaluation of an object by applying the map (Figure 2.1(c) and 2.1(d)) on a given input image (Figure 2.1(a)). Originally, this concept for *Visual Attention* was introduced by Itti, Koch, and Niebur [IKN98] who presented

a fast approach for scene analysis. They applied Saliency Maps that use biologically inspired complementary colors like red-green, blue-yellow, and intensity. In the context of object learning for robots features like, e.g., color, edges, and shape have proven to be best as they are the most important ones for humans according to [Sch01, Tso05]. After calculating Attention Maps they are usually fused to get a more reliable signal. Especially in the field of robotic research the required maps are often designed as neural networks, as they are assumed to appropriately simulate human behavior.

An applied example for an integrated approach using Saliency Maps for learning objects, directing the attention of a robot and selecting a Region-of-Interest is described by Gonçalves and colleagues. They have developed a framework for Robot Cognition [GWOG99]. In particular, they use a stereo head robotic platform for the processing of visual information. In their framework this information is used as basis for neuro-physiologically inspired Saliency Maps, similar to the proposed model for attention of Itti et al., e.g., [IKN98]. As a result Gonçalves et al. are able to create an environmental map containing pattern orientation, position, and representation. The collected information is then dynamically updated which enables adaptation to changes in the environment over time. However, their approach does not allow a human user to directly interfere which is one main aspect of a convenient Human-Robot Interaction as they mostly focus on the signal processing.

Besides this low-level signal processing with Attention Maps a cognitive motivated Personal Robot needs to provide high level cognitive functions like social skills as well. It is obvious that in a cooperative task, like playing a table game, the interaction will be more comfortable for the human user as long as the Object Attention System supports a natural interaction. This means that it has to support appropriate information about objects that can easily be processed by a deliberative dialog component. For instance, the Object Attention System needs to be able to establish connections to formerly learned objects and their symbolic descriptions, like color names and numeric color values. In order to prove that humans desire a robot that exhibits social skills they are accustomed to, Breazeal et al. [BBG+04] presented results of an investigation on humanoid robots that had social abilities. They state that it is imperative for a socially intelligent, cooperative humanoid robot to adhere people's *Social Model* in order to be able to predict, explain, and understand the robot's behavior. This Social Model assumes that humans automatically attribute different mental states (e.g., beliefs, feelings, ...) to non-living entities when they exceed a certain state of complexity. A Personal Robot that is intended to support untrained people in a natural and intuitive manner belongs to this category. Due to the increasing complexity of the interaction scenario the robot's social sophistication has to scale appropriately as well [FND03].

In order to face these challenges the next sections point out different theories of attention that are helpful for an interaction scenario with a Personal Robot. Starting with Joint Attention as a theory with the loosest coupling between mind and action (cf. section 1) over the mechanisms of Shared Attention up to the Theory Of Mind as the most sophisticated model as far as the development of the Object Attention System is concerned.

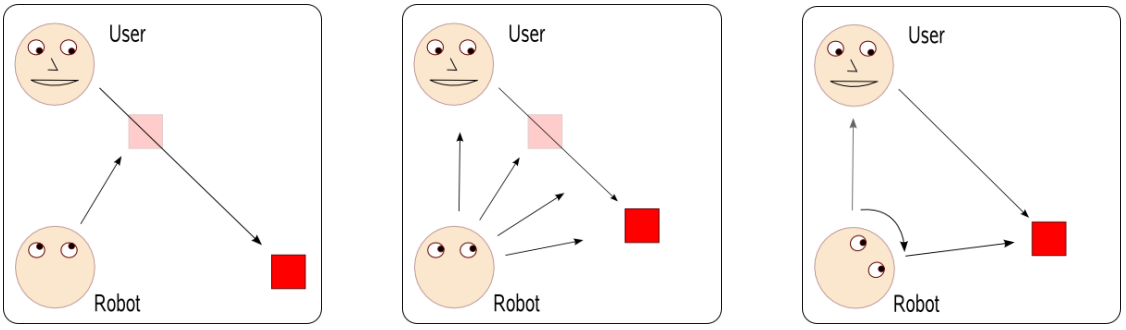
2.1 Theories of Attention

For an intuitive Human-Robot Interaction, the mechanism which directs the robot's attention to a referenced object can be realized more or less high-level. In the following, the most appropriate theories that are relevant for the development of the Object Attention System are presented from the perspective of increasing demands for cognitive abilities.

Joint Attention

Joint Attention as an advanced social skill for robots implies that the robot is able to follow the focus of attention of its interaction partner. In particular, Joint Attention offers a great opportunity for cooperative learning since the user is able to give direct feedback to the robot.

An approach that presents a model for the development of *Joint Visual Attention* in a robotic context is given in the work of Y. Nagai et al. [NHMA03, Nag04]. Subsequently, this work is exemplary used to explain the mechanisms of Joint Attention. Y. Nagai and colleagues motivate their learning attention model by the cognitive developmental process of infants. From this perspective, they simulate the staged learning process of infants in their behavioral model as it is introduced in [BJ91]. Next, this model is described by means of an interaction between a human user and a robot. In the first stage, which represents the *Ecological Stage*, the robot does not explicitly follow the gaze of the human user but tends to look at interesting objects that are within the gaze direction (Figure 2.2(a)). Compared to infants this behavior is usually shown at an age of 6 to 9 months.



(a) Ecological Stage: Robot looks at salient objects within the gaze direction of the user.

(b) Geometric Stage: Joint Attention only when object is observed in the field of view of the robot.

(c) Representational Stage: Joint Attention even if the object is out of the robot's current field of view.

Figure 2.2: Staged learning of Joint Attention.
The images have been adapted from [NHMA03].

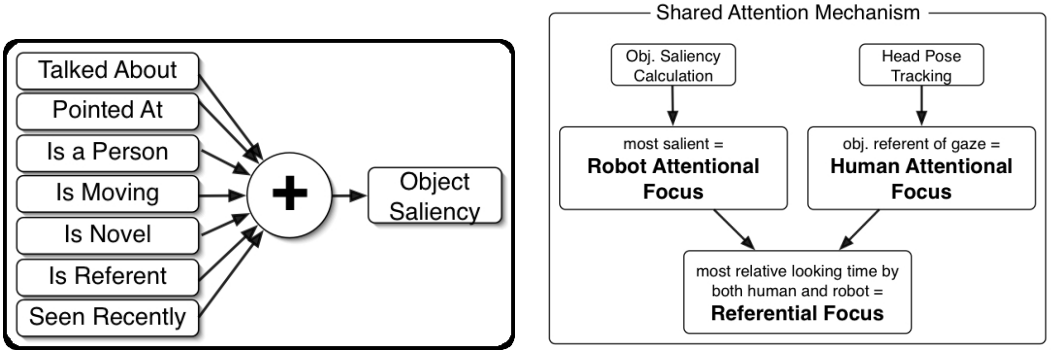
During the following *Geometric Stage* (Figure 2.2(b)), the robot realizes Joint Attention for the first time, but only if the user is looking at an object that is already within the robot's field of view. This stage of perception has been observed at infants at the age of 12 months. The *Representational Stage* shown in Figure 2.2(c) describes the last stage of development with regard to Joint Attention. The robot is able to follow the gaze of its user even if the Region-Of-Interest is initially not in the robot's field of perception. This behavior corresponds to experiments that

were done with infants at 18 months of age. All considered, a robot that should be able to focus its attention on an object referenced by the user, needs to simulate the behavior of an approximately 18 months old infant in order to enable an interaction humans are accustomed to.

Another approach investigating the development of Joint Attention as social cognition between a human user and a robot is proposed by Kaplan and Hafner in [KH04]. They state that besides viewing the gaze and simultaneous looking at a salient feature, Joint Attention also "...implies viewing the behavior of other agents as intentionally-driven..."[1]. Concluding, they propose that the challenges in the development of social cognition and Joint Attention, respectively, need to consider the coordinated development of intentional matching and inferencing, behavioral parsing and other skills. A step in that direction is given by models for Shared Attention which are described next.

Shared Attention

In this work the term of Shared Attention is used as extension of Joint Attention (cf. section 1). The main difference consists in the additional aspect that the involved interaction partners are aware of each other and know that they are referring to the same object. For instance, a model for a Shared Attention mechanism is presented by Lockerd Thomaz et al. [ALTB05]. For their studies they have realized their model for Shared Attention in their robot *Leonardo*.



(a) Influence factors for the computation of object saliency.

(b) Schematic of the Shared Attention mechanism used in the robot Leonardo.

Figure 2.3: A concept for Shared Attention with object context in an HRI. Explanations are given in the text. The images have been taken from [ALTB05].

Thus, Leonardo is able to direct its attention to the same object that the user refers to. As for the development of the Object Attention System the appraisal of objects is of major interest, the computation of the object saliency is discussed next. For its computation several aspects are considered in order to establish a social interaction scenario, where the overall saliency determination is the result of a weighted sum of different factors distributed on three categories. The first category considers the social reference (e.g., if something is pointed to) as it is depicted in Figure 2.3(a). The second category contributes the perceptual properties of an object, i.e., whether it's moving or its color. The third category concerns

[1] taken from [KH04]

the internal state of the robot. Depending on the current state the robot is, for instance, able to determine whether the object shown is familiar to the robot or not. After the object saliency has been computed, it is used in the overall attention model shown in Figure 2.3(b).

The attention model in turn is realized as an explicitly represented mental state in the style of the attention mechanism proposed by Baron-Cohen [BC91]. Within this state the model of Lockerd Thomaz et al. collects data that helps to determine what the person's interest is about. As the schematic of the Shared Attention model in Figure 2.3(b) illustrates, the model is divided into three components, in particular, three different kinds of foci. The first one deals with the current attentional focus related to the view of the robot. In order to be able to know what the human user is currently looking at, a second separate attentional mechanism for the human is used. These two distinct foci are combined in the third referential focus that involves the current topic of shared focus. Here, generic aspects are considered, like the topic of communication or performed activities.

Now that the principle mechanisms of Joint Attention and Shared Attention have been explained, they can be used as prerequisite in modeling a Theory of Mind, which is described next.

Theory of Mind

The different attention models described in the preceding sections enable the robot to direct its attention on the object of interest. But for a natural communication this is not sufficient as humans automatically attribute certain intentions, beliefs, or goals of the partner related to the object. For instance, if a person points to an apple, the person probably wants to eat it. This kind of inference can be described by the Theory of Mind which represents another essential part for a human-like interaction between a human and a robot. It is obvious that a Theory of Mind in a socially reactive robot provides several advantages. Aspects like, e.g., desires or emotions of the user could be realized in the robot's mind and, thus, it can react more appropriate. As an example, if the user is angry, the robot could try to calm him down.

An attempt to simulate this "mind-reading" behavior by integrating a Theory of Mind in a robot's architecture is described by Ono and Imai in [OI00] or Scassellati in [Sca02]. For example, Scassellati proposes a model for the humanoid robot *Cog*. In his work he adapted the models of Theory of Mind by Leslie [Les94] and Baron-Cohen [BC91]. Summarized, Leslie proposes in [Les94] that three classes of events based on their causal structure should represent the world. In brief, the *Actional Agency Class* explains goals and intents of agents, the *Attitudinal Agency Class* models attitudes and beliefs of agents, and the *Mechanical Agency Class* explains the rules for mechanics. In Leslie's model especially the interdependency between these classes is considered.

The model of Baron-Cohen [BC91] introduces a so-called *Mindreading System* that acts as a set of precursors for a Theory of Mind. His proposed system is shown in Figure 2.4 which is decomposed into four distinct modules. The first two modules, the *Intentionality Detector* and the *Eye Direction Detector* represent the

input that is provided by two perceptional representations. All stimuli with self-propulsion and direction (auditory, tactile, visual) on the one hand and all visual stimuli with eye-like shape on the other hand.

In detail, the Intentionality Detector creates a dyadic output for basic movements as well as declarations about approach and avoidance, like "She wants to eat the fruit." The output of the Eye Direction Detector determines whether the interaction partner is looking at the robot or not. Based on this information the module can interpret, e.g., the eye direction in order to guess the gaze and, thus, produces triadic representations like "Axel is looking (I see the apple)" while the three involved elements are “I", “Axel", and “apple". This structure is caused by the embedding of the first dyadic representation within the second one. Concluding, this means that both interaction partners attend to the same object (cf. section 2.1). The fourth module of Baron-Cohen’s Mindreading System consists of the Theory Of Mind mechanism. It connects epistemic mental states in other agents with our knowledge of mental states in one theory. So, intentions, desires, and beliefs of different agents can be predicted. Scassellati now links Leslie’s model together with the model of Baron-Cohen into a robotic Theory Of Mind for the robot Cog which is subsequently described.

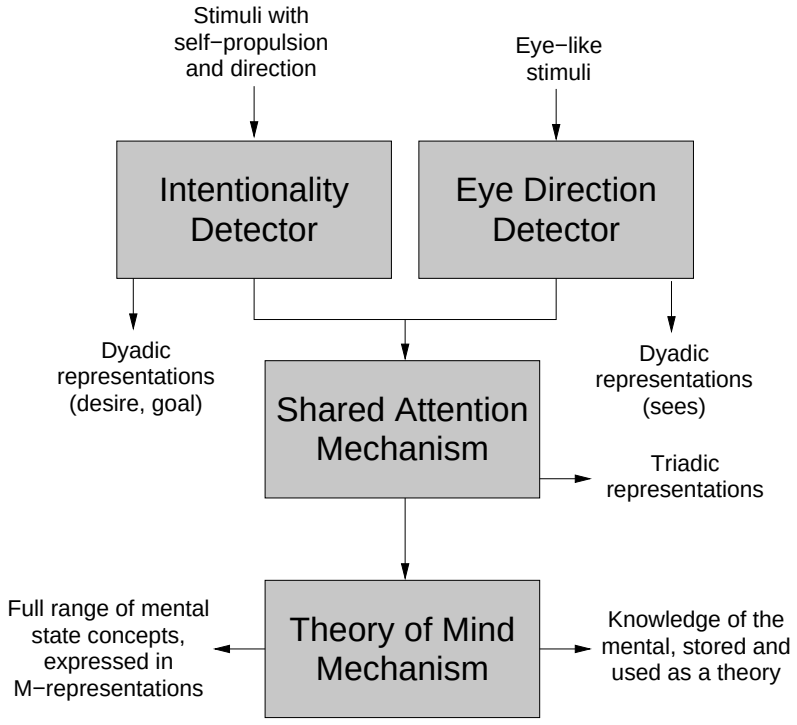


Figure 2.4: Baron-Cohen’s model visualized as block diagram for the development of a Theory of Mind in a robot. See text for description. Adapted from [BC95].

The initial system implemented by Scassellati in [Sca02] focuses on two abilities. First, it is able to distinguish between animate and inanimate motion and second it is able to identify gaze direction. For these tasks the system uses different detectors for Color, Skin, and Motion which are combined together with a habituational mechanism within an attention process. This attention model is based upon models of a human visual search and attention where this solution is held very flexible. Thus, it was possible to use a former implementation of this attentional mechanism on the social robot *KISMET* [BS99]. By using different input cues, the system is able to detect faces and follow the gaze in order to focus on the same object that a user attends to. An extension by imperative and declarative pointing of the presented system is described in [Sca03]. In that paper Scassellati and colleagues discuss their implementation of a behavioral model of social develop-

ment. In particular, they review the capabilities of the implemented mechanisms for Joint Attention. It has to be noted that they do not distinguish between Joint and Shared Attention. Thus, the aspects they report as Joint Attention are understood as Shared Attention in this thesis (cf. section 1). Although they reported an unfinished implementation of Baron-Cohen's model realized in the humanoid robot Cog the results gave promising insights. In detail, the Eye Direction Detector and the Shared Attention module as well as a mechanism to orient the neck and arm pointing were used. Based on this configuration, the users who interacted with the robot found its social behavior, for instance, head-nod imitation and eye-neck orientation both believable and entertaining. Summarizing, it can concluded that the implementation of a Theory Of Mind is a step in the right direction during the development of cognitive Personal Robots and the proposed Object Attention System, respectively.

However, the mental models of attention present only one side of the coin, the other one copes with the integration of perceptual sensors in a mentally motivated design. For this reason, the next sections will focus on architectures used in social robots that offer Object Attention capabilities. Especially as such architectures provide an infrastructure that supports an assignment of sensor signals to symbolic expressions which is known under the term *Anchoring*, cf. [CS03, FKL+03, Lan05].

2.2 Architectural Aspects for Robotic Systems with Object Awareness

The control of attention can eventually be realized based on one modality only as input cue. Especially for mobile robots where computational resources are strongly limited, unimodal solutions are advantageous because they usually have a reduced consumption of computational power since less data has to be processed. The disadvantage of unimodal processing lies in the vulnerability for perceptional errors. This problem can partially be compensated by multimodal approaches using more than one kind of input, e.g., vision and audio.

To give an overview of different approaches, the next sections will review attentional mechanisms with a focus on unimodal and multimodal input processing for their advantages and disadvantages in robotic systems that provide a kind of Object Attention.

2.2.1 Unimodal Attention Processing

Unimodal attention processing for socially reactive robots is not very common since a natural *Human-Human Interaction* usually uses several modalities (e.g., Speech, Gestures) as input. Like described above, unimodal solutions are very sensitive to errors and, thus, a selected modality often does not provide reliable data. For instance, if in a purely vision-based system the Region-Of-Interest is temporarily occluded, it's very difficult to enable or uphold the focus of attention on the referenced location. As a second example, a purely audio-based approach will fail if the background noise (e.g., talking people in the robot's vicinity) overlays the speech of the *Person-Of-Interest* and as a consequence the needed information

can not be extracted. For a purely audio-based Object Awareness approach the objects always need to produce a sound. But since sonar-based or ultrasonic approaches are too coarse in order to pinpoint and identify a far-off object they are of minor relevance for a natural interaction as well. Furthermore, systems using wireless identification, like radio frequency tags, are not the focus of this work either, as they are not used in a natural interaction. Thus, only architectures that are mainly vision-based will be discussed subsequently.

This section is structured as follows. First, a general approach of establishing Joint Visual Attention that allows to focus on an object is presented. Subsequently, a model that enables a coarse fixation on an object in a large cluttered scene is described. However, on the contrary to a large cluttered scene interactions often take place in a small restricted area with similar looking objects. Thus, a proposal that allows to focus the visual attention on a single object in a strongly restricted scene with a couple of similar objects standing nearby to each other is discussed.

The Joint Visual Attention approach by Y. Nagai et al. [Nag05a, Nag05b] (cf. section 2.1) integrates a computational learning model that is used to comprehend the development of Joint Visual Attention by investigating pointing, reaching, and tapping gestures. In the following, their learning model is explained on basis of Figure 2.5 that illustrates the overall setup.

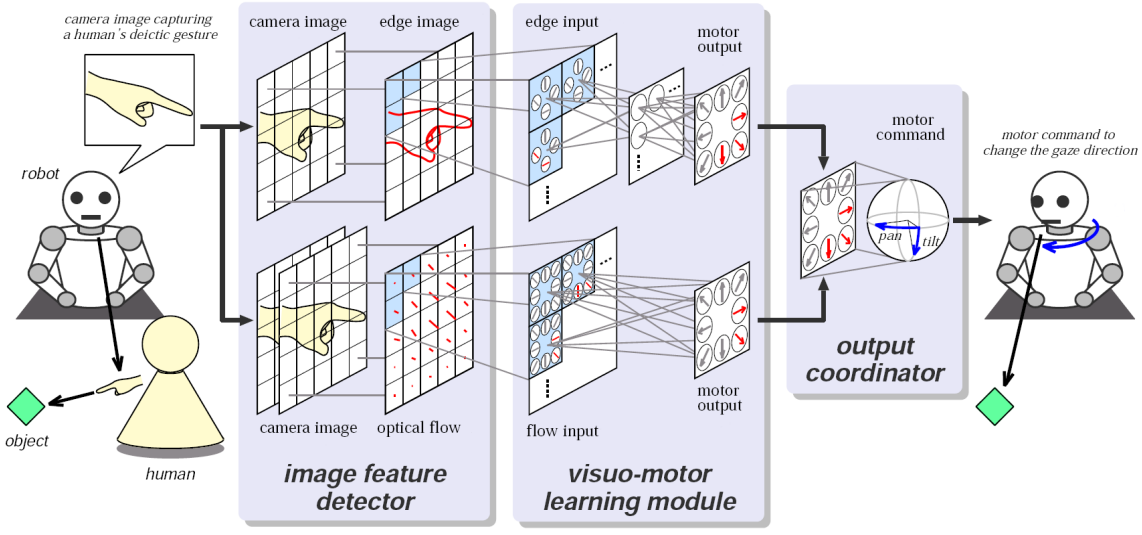


Figure 2.5: Learning model for the acquisition of Joint Visual Attention. See text for a detailed description. This figure has been taken from [Nag05a].

As experimental input for the setup, a human user performs the above mentioned different types of gestures (pointing, reaching, tapping) which are evaluated by a salient *image feature detector* that analyzes the edges and the optical flow in the captured images. Thus, the movement of the hand can be determined. To do so, they apply a *visuo-motor learning module* that uses two neural networks. As a consequence it becomes possible to let the robot follow the pointing direction in order to determine the Region-Of-Interest which contains the salient object. The execution of the calculated following behavior is realized in a last processing step. Here, the output of the feature detector and the learning module is verified by the

Output Coordinator that decides which motor commands are actually sent to the robot's head.

The results of the conducted experiments showed that the chosen category of a gesture affects the learning process regarding the comprehension of the direction of a gesture. Y. Nagai comes to the conclusion that the reaching gesture facilitates the fastest learning method to comprehend the direction of a gesture because of the richness of edge features and the closeness to the referenced object. The second fastest learning results are possible with the tapping movement as it offers a qualitative difference between the edge directions and the optical flow. The remaining third category of pointing gestures are not that suitable due to their distance to the referenced object as this involves the capability to correctly assign an object that is only within the pointing direction and, thus, more difficult to determine.

Concluding the experiments of Y. Nagai, the best possible results for the determination of the Region-of-Interest by the proposed Object Attention System can be expected if the user directly touches the referenced object. However, as it cannot be assumed that the object is always nearby the user so that he can directly touch the object, a coordination of the robot's head might become necessary in order to roughly focus on the area containing objects, see Figure 2.6.

Figure 2.6: Focused object (sign with red boundary) after head-eye saccade, taken from [VCSS01].

Since these difficulties require an active mechanism to direct a "spotlight" of attention, Vijayakumar et al. investigated in [VCSS01] the computational mechanisms for Visual Attention. For this purpose they have build a biologically inspired system (cf. [SF03, KLT04]). In particular, they modeled an artificial oculomotor system on an anthropomorphic robot, like it is shown in Figure 2.7. The schematic illustrates that the necessary computations are modularized into three distinct subparts, the *Sensory Processing Module*, the *Motor Planning Module*, and the *Interaction Issue Module*. The Sensory Processing Module mainly consists of a *Competitive Dynamical Neural Network* for modeling cortical information processing. It is responsible for the conversion between the raw sensory input signals and the camera coordinates that are used as target for the next saccade.

Then, the Motor Planning Module takes these coordinates in order to transform them into motor control commands for the head and the oculomotor system. Last but not least, the Interaction Issue Module is used to control higher level actions of overt attention. This involves for example the cancellation of self-motion as a potential target of attention. To prove the usability of the developed system, Vijayakumar et al. demonstrate the performance of their attention model by focusing on different small objects (round signs) in a cluttered scene, cf. Figure 2.6. Unfortunately, a natural environment usually offers a variety of different objects standing nearby to each other. Thus, such a coarse visual search, like it is presented by

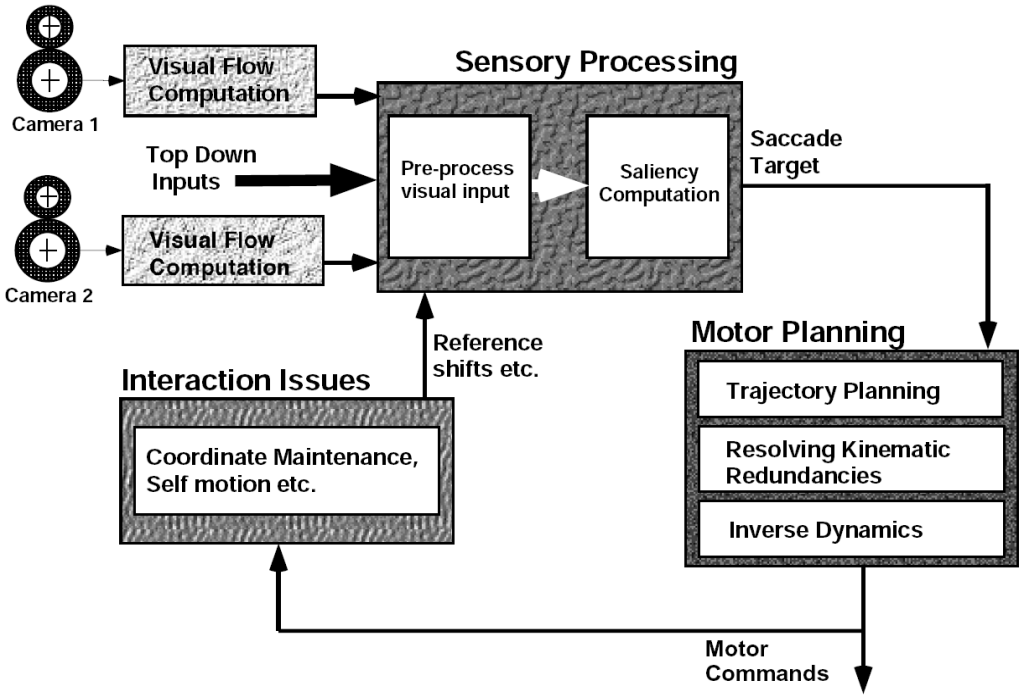


Figure 2.7: Schematic of a system implementing overt Visual Attention. See text for a detailed description. This figure has been taken from [VCSS01].

Vijayakumar and colleagues is often not sufficient for the use in the proposed Object Attention System. An approach that allows a closer visual search within the scene is, therefore, described next.

A few approaches, e.g., [BE04, KBCE05, HW06a] deal with aspects of directed attention for objects within a cluttered scene. For instance, in [HW06a], Hawes and Wyatt present an extension of the Itti & Koch model of Visual Attention by contextual information. Their approach is developed as part of the EU Project *Cognitive Systems for Cognitive Assistants (CoSy)* which defines the so-called *PlayMate scenario*. Within this scenario, a human and a cognitive inspired robot interact with objects placed on a tabletop. To improve the scene analysis of the tabletop view, Hawes and Wyatt altered the approach of Itti et al. [IK00] which is depicted in Figure 2.8. The difference is not to linearly combine the computed Conspicuity Maps for colors (red-green, blue-yellow), orientations (0°,45°, ...), and intensities and then to fuse them in a single Conspicuity Map

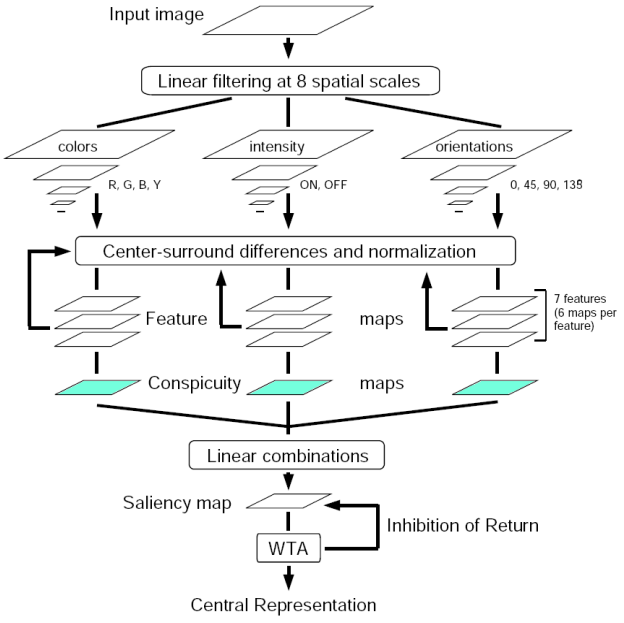


Figure 2.8: Schematics of Itti & Koch model for Visual Attention, taken from [IK00].

(highlighted layer in Figure 2.8), but to integrate the different single Conspicuity Maps directly into the final Saliency Map.

In the end, this final map is used as input for a *Winner-Take-All (WTA)* neural network that detects the most salient location and, subsequently, directs the focus of attention towards it. For context-sensitivity they additionally use weights in the final linear combination in order to enable varying presence in the Saliency Map. As a first evaluation result they report improvements in performance towards the original Itti et al. model on a static scene with a number of different-colored cans on a white tabletop. This result is very interesting with regard to the application in the Object Attention System as an increased performance is helpful to support a convenient Human-Robot Interaction. Thus, it has to be considered, whether a similar approach can be realized in the Object Attention System as well.

Summarizing this section, it turned out that unimodal approaches are only suitable to investigate basic behaviors, like, e.g., the development of different kinds of attention. For an intuitive Human-Robot Interaction, however, at least Speech or Gestures are required. Thus, a couple of multimodal architectures using these input cues are discussed next.

2.2.2 Multimodal Attention Processing

Multimodal attention models that are using different input cues to focus the direction of attention, underly the same principles as unimodal approaches. But due to a combination of different modalities it is possible to compensate the weaknesses of a single input cue up to a certain degree. Here, integration in a spatio-temporal sense becomes one of the most important issues. Relying only on one modality has been proven as insufficient input for directing a robot's focus of attention in domestic domains which represent dynamic, cluttered, and noisy environments. In particular, variations in lighting conditions or moving and talking people usually cause at least one sensor not to support a robust cognition of the environment. These variations demand for a flexible solution while developing the Object Attention System. Additionally, due to the unstructured environment, the robot always needs to be able to learn new objects and locations. In [RP97], Roy and Pentland propose that only multimodal interfaces are able to provide a solution for these challenges. They point out that it requires at least two different modalities to learn a new circumstance, e.g., the meaning of a new object. First, the main input cue that is responsible for the "task" itself (here, a new object) and, secondly, an additional source of information (e.g., a gesture) which indicates that an object should be learned.

The outline of this section is as follows. This section begins with biologically inspired robotic architectures of increasing complexity used in two different robots. The first one facilitates a cognitive motivated architecture which uses speech for the description of objects. The second example discusses an architecture that, furthermore, offers a Theory Of Mind. Both examples have in common that they have been applied on stationary robots. To point out that similar approaches are applicable on mobile robot platforms as well, two mobile robots are introduced that, in particular, support speech and gesture processing. The latter one, additionally, does a visual object analysis.

One example for a stationary robot providing an Object Attention mechanism is the Bielefeld robotic system *GRAVIS* which is described in [SRHR04]. The GRAVIS system directs its attention towards an object while the user is pointing to it. The diagram in Figure 2.9 depicts its principal architecture.

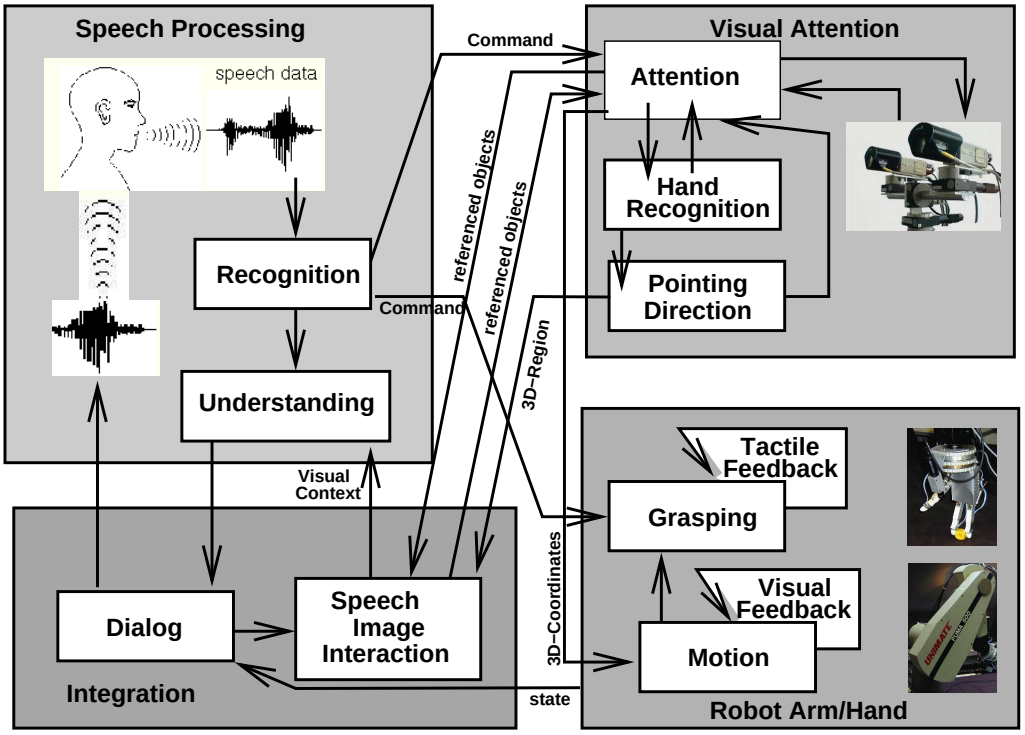


Figure 2.9: The Bielefeld robotic system GRAVIS.
See text for a detailed description. Image taken from [SRHR04].

Basically, it is divided into four modules, the *Speech Processing*, the *Visual Attention Processing*, the *Control Mechanism* for the arm and hand manipulator, and the *Integration Module* responsible for the convergence of linguistic and visual/gestural inputs. All considered, these modules are able to generate saccadic movements of the stereo camera head in order to support an active scene analysis for finding and interacting with objects in the scene. This method of Object Attention is supported by the layered architecture (Figure 2.10) of the Visual Attention module. It integrates several Feature Maps. Specifically, the different maps highlight the saturation and intensity within the **H**ue **S**aturation **I**ntensity *(HSI)* color space. Thus, skin color, oriented edges, and motion based on a difference map can

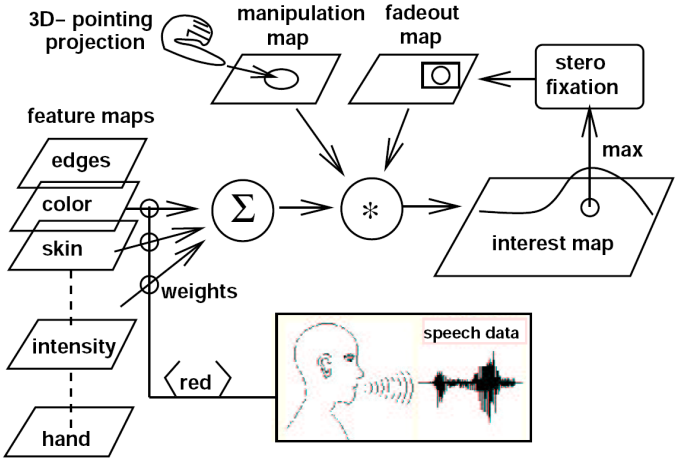


Figure 2.10: Attentional subsystem of the GRAVIS robot, see text for a detailed description. The image has been taken from [MFR+02].

be highlighted. Spoken simplified, a "moving skin map" is created as an additional Feature Map. Furthermore, a map for the detection of deictic pointing gestures is added. During the subsequent processing a weighted sum of all Feature Maps is combined with a manipulation map and a fadeout map in order to create a final interest map. Finally, this map is overlayed with a Gaussian smoothing to suppress small saccades and, consequently, the dominant peak is used as new fixation point. Hence, a stereo matching is activated and the resulting loop continuously generates new saccades.

Current research of this architecture deals with a mosaic image representation used as a Short-Term Scene Memory. Besides, the current research uses a new robotics platform representing the successor of the GRAVIS system which now integrates a 20 Degrees-of-Freedom (in the following denoted as DoF) Shadow Dextrous Robot Hand.

To sum up, the GRAVIS robot is able to pinpoint a location of an object and, therefore, after an object formerly has been learned it is able to grasp and interact with it. Regarding the support for Object Attention in the GRAVIS robot, its architecture does not allow to learn verbally specified additional object properties, as the work mainly focuses on imitation grasping. Hence, it definitely provides an excellent model to resolve object references, but it limits the knowledge stored about objects.

An example for an architecture supporting Object Attention which is integrated in a very sophisticated stationary robot is given next. Breazeal et al. describe in [BBG+04] the architecture shown in Figure 2.11 that is used in the cognitive robot Leonardo. Similar to the architecture of the GRAVIS robot it is divided into several subsystems.

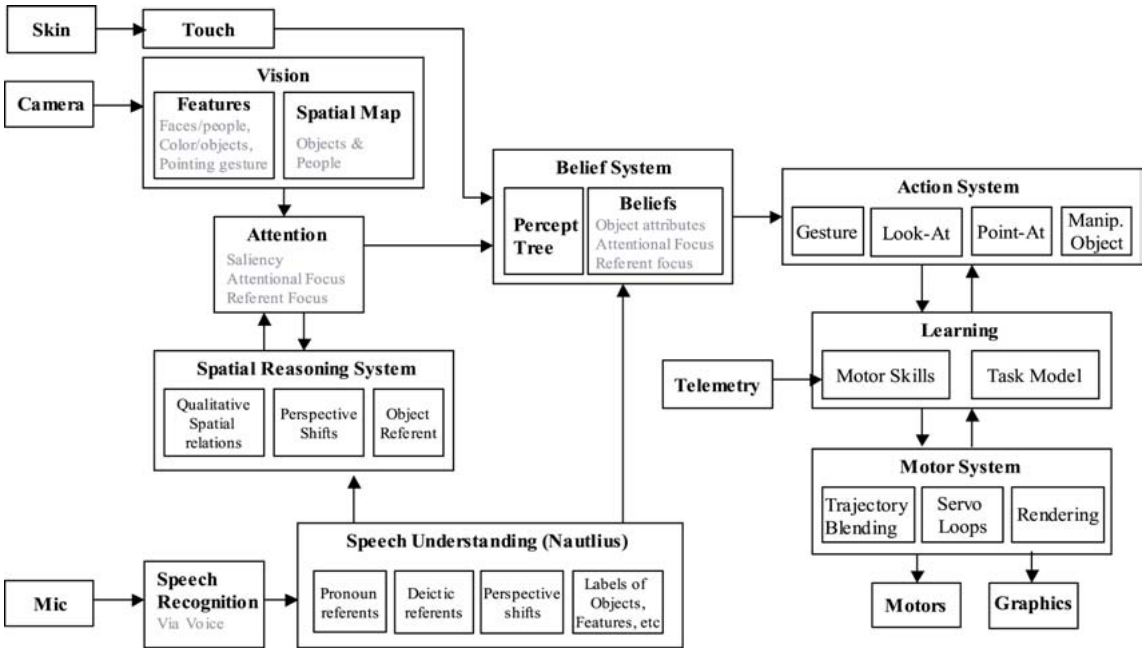


Figure 2.11: Architecture of the cognitive robot Leonardo.
See text for a detailed description. Image taken from [BBG+04].

Beginning in the lower left of Figure 2.11, Leonardo has a *Speech Processing Unit* including the Speech Recognition and Speech Understanding. Besides, the robot

offers a *Spatial Reasoning System* which acts as a mediator between the Speech Processing and the Vision Processing which in turn includes a Shared Attention mechanism, cf. section 2.1 on page 13. Both, the Speech Processing Unit and the Vision Processing Unit support a *Belief System* that enables Leonardo to reason about a referenced object. Thus, a common focus of the human and the robot on the same object becomes possible. This is enabled by the evaluation of vision and speech, e.g., the user is pressing a red button and, simultaneously, saying "That's the red button". The *Action System* which supports a *Learning Module* addresses the *Motor System* for appropriate actions of the robot.

Summarizing, the robot Leonardo is able to learn new objects including verbally given properties. Although this matches the aim of a natural Human-Robot Interaction, the robot depends on a well-defined environment with special cameras observing a particular part of the scene. For instance, a camera mounted at the ceiling for the observation of the restricted object area in front of the robot. To overcome these limitations of stationary robotic setups, approaches integrated on mobile robots have been developed which are presented next.

Within the last ten years more and more robotic architectures considering visual object appearances, speech, and gestures have been developed. In order to name few of them, e.g., the Perseus architecture [Kah96], the architecture for the robot *HERMES* [BG99], the service robot *Albert* [REZ+02], or the humanoid robot *ARMAR* [BSZD06][2]. To get a better insight of the architectures used in such multimodal operating robots, in the following we concentrate on two canonical examples. Therefore, the approaches of Ghidary et al. [GNS+02] and Kruijff et al. [KKH06] are subsequently described.

Ghidary and colleagues use for their mobile robot the architecture shown in Figure 2.12. As it can been seen, it is divided into two main parts. First, the complete speech processing is done on a separate host computer, and, secondly, everything else, e.g., visual processing or behavior-control is processed directly on the robot. The mobile robot of Ghidary et al. is used for room map generation while it is able to learn rough squared 2D-views of objects supported by static hand postures and predefined speech commands given by the user. In order to control the robot, Ghidary et al. use the so-called *Behavior Controller*, represented by a *Finite State Machine (FSM)*. As main control mechanism it determines the most appropriate behavior sequence in order to reach a particular goal.

For the learning and localization of new objects, Ghidary et al. use a depth-from-focus approach supported by an autofocus camera. This enables the robot to measure the distance to the hand and the object, respectively. It has to be mentioned that their approach does not support an object recognition, instead they estimate an object's position based on the position of the user's hand. In particular, the current hand position is focused and, subsequently, centered by the result of the skin color detection using the center of mass of the segmented hand. As the environment and the robot is equipped with a Home Robot Positioning System (HRPS), the robot is, thus, able to enter the exact position of the hand and the object/location where the user pointed to, respectively, with a squared image pattern in a knowledge base. This knowledge base is used to generate

[2]Part of the Collaborative Research Center 588 "Humanoid Robots – Learning and Cooperating Multimodal Robots" and the German Service Robotics Initiative DESIRE [fPuA05].

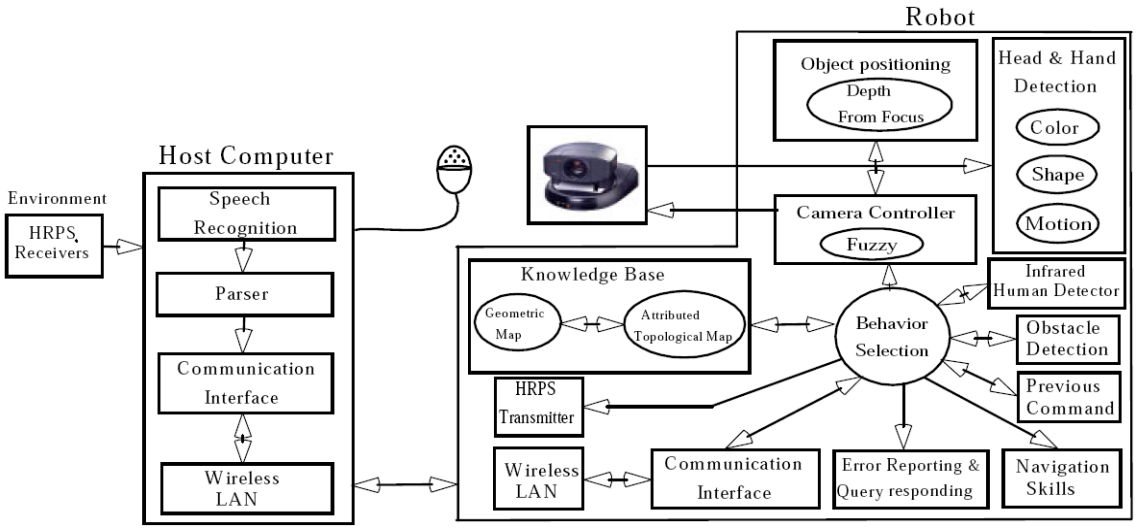


Figure 2.12: Control Architecture of the mobile robot used by Ghidary et al. See text for a detailed description. The image has been taken from [GNS+02].

a map of the robot's environment including all referenced objects with their relations to each other and their verbally specified sizes. Once an object view and its location has been learned, this information can be referred in the knowledge base with corresponding queries. Thus, the map can be updated over time as in, e.g., a domestic domain the position of objects (e.g., dishes) often changes. The approach of Ghidary and colleagues, however, mainly focuses on the detection of the human user, and, hence, only a basic model of Object Attention is implemented. Unfortunately, the knowledge that can be gained about objects is very limited (size, estimated position, relation to other objects). Furthermore, due to the use of the indoor positioning system using infrared and ultrasonic sensors, the robot can construct an adequate map only in especially prepared environments which conflicts with the requirements of a natural surrounding. A cognitive model that overcomes these limitations and provides a more extensive mechanism for Object Awareness is discussed next.

Kruijff et al. currently use two different modalities for the realization of an Object Attention mechanism [KKH06]. They state that "...reference resolution in a situated dialog is a particular instance of the anchoring problem [CS03]..."[3] related to the correspondence of sensor data and symbols that refer to the same physical object, cf. J. Fritsch et al. [FKL+03]. In order to realize a solution for the anchoring problem, they integrate a combination of two different fusion strategies, an intra-modal one versus an inter-modal strategy. The intra-modal strategy generates different object hypotheses occurring in a single modality which concern one and the same object. As an outcome, equivalence classes are generated which store the information about different occurrences in a conceptual manner. In order to define relations between the classes that were created in different modalities, the inter-modal strategy is used. Here, the approach of Kruijff et al. uses an ontology-based mediation which utilizes the conceptional information gathered in each equivalence class. To prove the efficiency of the presented strategy, a

[3] taken from [KKH06].

framework for the resolution of object references occurring in a Human-Robot Interaction has been developed, see Figure 2.13.

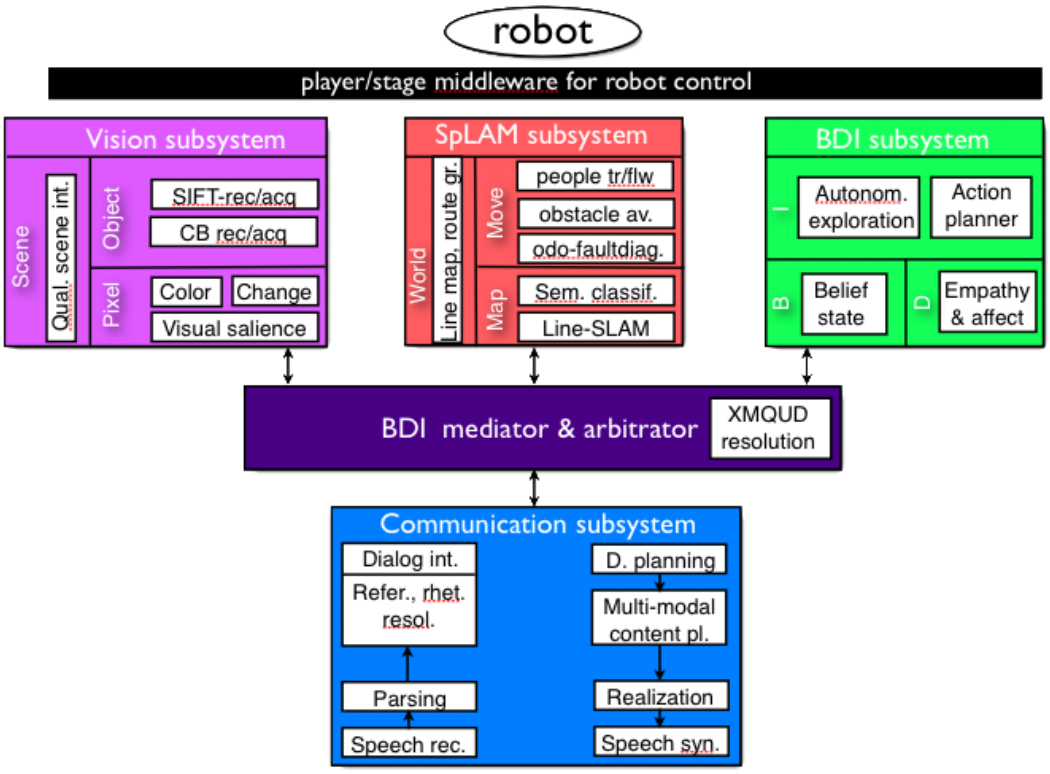


Figure 2.13: Control Architecture of the CoSy project.
See text for a detailed description. The image has been taken from [KKH06].

The figure illustrates that the architecture is realized as a three-layered hybrid architecture, cf. J. Fritsch et al. [FKH+05], based on a **B**elief-**D**esire-**I**ntention *(BDI)* process to mediate between different subsystems. The three subsystems developed are responsible for communication issues (blue highlighted), spatial localization & mapping (red highlighted), and visual processing (pink highlighted). Now, to achieve a common ground between the different modalities, the BDI process uses beliefs.

The visual processing responsible for the Object Attention uses three cues, identity, color, and the size of objects in a scene [KKH06]. As the diagram in Figure 2.14 shows, SIFT features [Low04] are used for the identity computation. The given example illustrates incremental learning of a visual object supported by corresponding symbolic speech input.

In case that an object which is unknown to the robot should be learned, the detailed processing is discussed next as this is a major issue for the development of the Object Attention System. Starting with an utterance like "This is a box", a new SIFT-based model should be learned. Therefore, in order to create an object's identity, Kruijff et al. first segment the object view by a bounding box with fixed size. Therefore, the object that should be segmented needs to be placed without occlusions on a white tabletop. Then, SIFT features are extracted and, subsequently, stored in a new equivalent class together with a description of the object (here: "box") [KKBL06]. In correspondence to the new equivalence class, the BDI

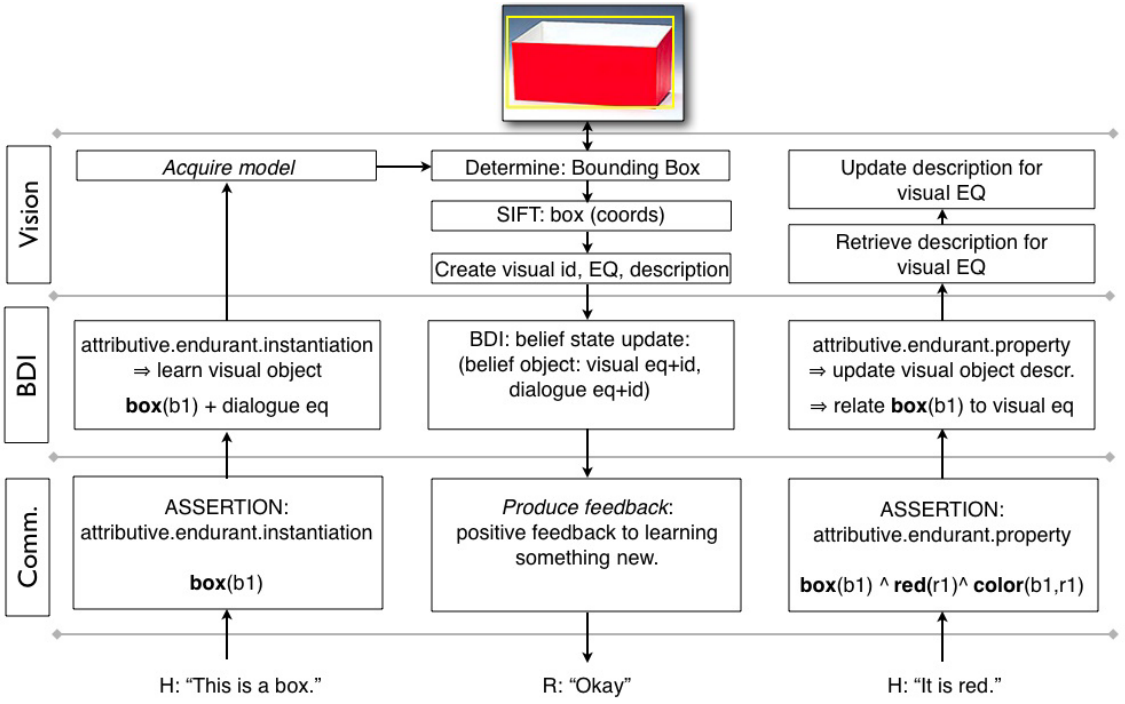


Figure 2.14: Diagram of multimodal object learning used within the CoSy project. See text for a detailed description. Image taken from [KKH06].

mediation creates an appropriate belief that enables the dialog to create a suitable feedback utterance for the user. If the user now says "It is red", the additional information for the color is linked by the keyword "it" to the current discourse referent for "box". Consequently, the mediator informs the vision subsystem about the new property "red" and updates the corresponding visual equivalence class. In particular, a histogram determines the color which is applied on the segmented region using the HSI color space. After the histogram is smoothed in a subsequent processing step, the dominant color peak indicates the object's color value.

In addition to the learning of previously unknown objects, the system is able to recognize already learned objects as well. For this task, SIFT features of the current scene were calculated. In case that the found number of features exceeds a certain threshold, the affinities between the already stored object features and the currently extracted features result in the computation of an affine transformation. Based on this transformation they determine the object's pose while it is applied to the model's segmentation mask. Although it is only possible to learn and to recognize a single object in a scene, Kelleher and Kruijff are able to establish relational descriptions between the objects learned within a qualitative scene representation [KK05]. Such a scene representation can be compared best with the Scene Model presented in this thesis, cf. the overview illustration 1.3 on page 7. This issue is discussed in detail in the corresponding section 5.1 on page 90.

All considered, the approach of Kruijff et al. is only partially usable for a natural scene, as, e.g., domestic domains are very cluttered and, thus, it is not realistic to assume that always a strictly separated object is present. However, a realistic environment analysis is not the focus of their experiments.

To sum up, a few recently developed multimodal approaches for Object Attention, partly biologically-inspired and designed for a cognitive Personal Robot exist. Nevertheless, a separated module for Object Attention within a mobile robot supporting well-defined flexible interfaces is a challenge neglected so far. Here, the proposed Object Attention System presented in this thesis enters a new domain, as it not implicitly integrated in an architecture but allows due to its modularity to be integrated in architectures that have either a very limited perception of objects or do not offer an object awareness at all.

A couple of robots have been developed as host systems for the different Object Attention mechanisms that have been presented in this section. Therefore, some of these different hardware platforms that either apply the above mentioned architectures or at least facilitate similar approaches that also allow to cope with objects during an interaction are introduced next.

2.3 Robots paying Attention to Objects

In this section, a variety of the most sophisticated humanoid robots that provide an Object Attention mechanism is presented. In accordance to their multimodal architectures which were partly described in the previous section, only robots using multiple input cues are considered as they enable the most natural Human-Robot Interactions.

Care-O-Bot II

The service robot *Care-O-Bot II* [HGS02] was primarily developed for interactions with elderly people. Thus, it offers a number of different sensors which enable an easy Human-Robot Interaction. Besides a 6 DoF manipulator arm used for object-related tasks, adjustable walking supporters, a tilting sensor head with two cameras, a laser range finder, as well as a control panel are integrated. With the help of these sensors, the robot supports some interesting interaction tasks.

In particular, the robot is able to learn, recognize, and grasp objects [GHS04]. Furthermore, a map of the environment with different landmarks can be processed which can basically be seen as a kind of Scene Model although the object locations are not explicitely denoted. This becomes possible through a remote control which enables the user to transmit commands to the robot. In order to learn a new object model, a camera captures an image of the unknown object and, subsequently, the view is stored in a database. In case the robot should grasp a known object, its image is retrieved from the database and compared with the current camera view. Additionally, a 3D-laser scanner provides distances to determine the exact object position. Thus, the robot

Figure 2.15: Care-O-Bot II. © Fraunhofer IPA.

is able to generate a trajectory for its arm and can grasp the object. Finally, the robot brings the object to the user.

Although Care-O-Bot II can interact with objects, the user always needs to use a control panel to instruct the robot. Hence, it is indeed a helpful service robot, but due to the missing speech and gesture processing, it matches only partially the flexibility of human-like interactions. A robotic system which is more interaction-oriented is the robot *HERMES*.

HERMES

Figure 2.16: HERMES.
Image taken from [BG99].

The impressive humanoid robot HERMES presented in [BG02] is a mobile robot as well. It has got two arms which have 6 DoF each and for direct manipulations of objects, 2-finger grippers are mounted at the end of the arms. Additionally, to optimize the position, e.g., while the robot puts down a tablet with objects on a tabletop, the upper body can bend forward and backward. For visual processing tasks concerning objects, HERMES has got 2 pan-tilt cameras mounted on a pan-tilt head that is connected to its torso. HERMES also has tactile sensors around its base and an auditory system integrated for object manipulations and speech recognition, respectively.

A long-term study of 6 months in the Heinz-Nixdorf museum in Paderborn in Germany enabled the robot to demonstrate its capabilities to a larger audience. In particular, HERMES is able to interact with objects in different ways. It can detect, recognize, and also track multiple objects. As it is, furthermore, able to build a map of the environment by a vision-based navigation system, the robot can drive to specific objects and locations. As soon as the robot arrives at the destination, the manipulator arms can be used to interact with objects, e.g., taking over a glass from a person. Finally, the robot can be instructed by naturally-spoken commands and even talk to interaction partners in the languages German, English, and French.

Summarizing, the robot HERMES is well-suited for, e.g., simple guidance or fetch-and-carry tasks, but it does not support higher cognitive functions, like, for instance, talking about already known objects or learning additional information about object features. Compared to Care-O-Bot II it is more suitable for a convenient interaction as it is equipped with an interface for naturally spoken language. However, as it does not support a gesture recognition, it still does not match all requirements of an intuitive interaction. A few more robots have been developed that support a natural Human-Robot Interaction. One example is the mobile robot *HOROS*, described in the next paragraph.

HOROS

The mobile *Home Robot System (HOROS)*, presented by Richarz et al. [RMSG06] is another cognitively motivated *Service Robot* that is from a mechanical point of

view very similar to the robot BIRON that has been used for the evaluation of the proposed Object Attention System. HOROS is equipped with a 180° laser range finder and three cameras for visual perception of the environment and the user. The omnidirectional camera is mounted on the top (Figure 2.17), another camera with a telephoto lens is mounted on a tilting unit at the front part of the head. Additionally, a camera with a wide-angle lens is located in one of the eyes. Furthermore, a pair of microphones is integrated into the platform as well as a touch-sensitive tablet PC. With help of these sensors, a speech, and a gesture recognition system, HOROS is able to detect interacting users and can navigate to Regions-Of-Interest that are referenced by a pointing gesture of the user. For navigational tasks, HOROS can access a given environment map that, however, does not include any information about specific objects, hence, no adequate Scene Model for objects is available, yet.

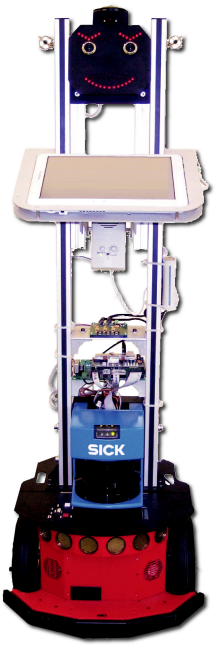


Figure 2.17: HOROS. Image taken from [GRM+06].

Although the robot does not have an explicit object-centered attention model yet, it is already able to resolve location references based on gestures which enables a mechanism for Object Attention as well. In this aspect it is less sophisticated than the robot HERMES but on the contrary it offers a gesture-based interface. A robot that is able to focus on a referenced Region-Of-Interest and recognize already known objects is realized with the robotic platform *Infanoid*.

Infanoid

Infanoid is a stationary infant-like robot [KY01] with 23 DoF, developed for investigations of the cognitive development shown by human infants. In order to be able to simulate their behavior, the robot is equipped with a stereo camera head with 3 DoF. With regard to the stereo head, each eye has got two built-in color cameras with different focal lengths (foveal and peripheral view) that offer, additionally, 2 DoF. Furthermore, its neck is equipped with 3 DoF as well. For manipulation tasks, Infanoid has got two arms that have 6 DoF each and a trunk that offers 3 DoF.

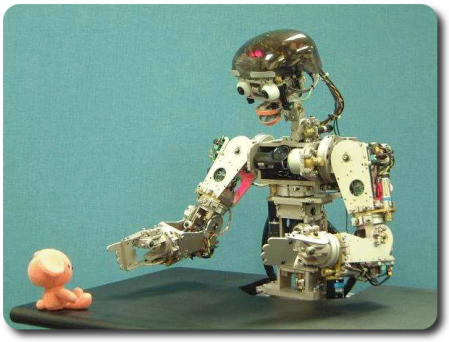


Figure 2.18: Infanoid. Image taken from [Nag04].

In accordance to the intended use as investigation platform for the development of Joint Attention in human infants, the robot has been used, e.g., in [Nag05b]. As described in section 2.2.1 on page 17, Y. Nagai et al. simulated the behavior of Joint Attention using gaze and gestures. In their approach, they used a purely vision-based analysis of gaze and gestures. Their model successfully described the possible role of gestures in the development of Joint Attention in infants. Nevertheless, in the context of modeling an Object Attention, the missing speech processing does only allow a restricted Human-Robot Interaction as the robot is only

able to follow the gaze or gesture direction of its user to focus on a salient-colored object. It does not allow to learn additional object attributes or other features that can verbally be specified, like a position. Furthermore, no learning of unknown object instances or the construction of a Scene Model is supported which, however, is essential for autonomous actions of the robot.

To sum up, the learning of unknown object instances is a challenging task. Therefore, only very few research groups cope with this topic. One example for such a robot that is able to learn unknown attributes, like color, about objects is the robot Leonardo (cf. section 2.1).

Leonardo

The robot Leonardo, depicted in Figure 2.19, is developed at the Massachusetts institute of technology as successor of the robots COG (cf. section 2.1 on page 14) and *KISMET*. Hence, its architecture and capabilities represent an advancement of the former work for an Object Attention mechanism.

Leonardo is a 65 DoF embodied robot with an expressive 24 DoF face, an active 4 DoF binocular vision system in the eyes, two actively steerable 3 DoF ears, a 4 DoF neck, two 6 DoF arms, and two 3 DoF hands with tactile sensation. Furthermore, Leonardo is a stationary robot using two additional vision systems, the first one is mounted behind the robot to provide a peripheral view for tracking people and objects. The second vision system is a stereo camera in the ceiling directed on the workspace in front of the robot. It is used to track objects and pointing gestures.

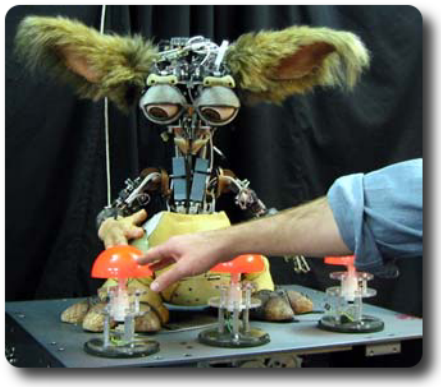


Figure 2.19: Leonardo.
Image taken from [BBG+04].

Leonardo is capable to detect the interaction of a human with saliently colored buttons arranged in front of the robot. To do so, Leonardo recognizes objects using deictic gestures in combination with speech and salient object features (cf. its architecture in section 2.2.2 on page 22). In particular, it is possible to label buttons by giving verbal information as well as to teach the robot how to use these buttons.

However, due to its stationary setup, it is not easily possible to transfer its Object Attention approach to other platforms as, e.g., the gesture recognition works from a totally different perspective than an ego-view-based recognition system does. Additionally, the robot is not able to construct a map of the environment that can be used as Scene Model.

To finish the overview about robots that have a human- or object-oriented attention mechanism, the robot *Robovie* is presented next as a final example.

Robovie

A couple of research groups, e.g., [NHMA03, MI05] have been used the robotic platform Robovie [IOI+01] (Figure 2.20) for the integration of an Attention System that is able to detect objects. For this task, the mobile robot is equipped with two arms (4 DoF each), a head (3 DoF) and two pan-tilt cameras serving as eyes. Additionally, it offers an omnidirectional vision sensor, 10 tactile sensors, two microphones, and 24 ultrasonic sensors.

An approach that deals with an integrated Attention System is, for instance, presented by Mukai and Imai [MI05]. They have implemented a communication system using Feature Drift in order to, spontaneously, change the attention to specific objects (colored blocks) in the vicinity of the robot. Besides the visual inspection of the scene, their approach contains a speech recognition unit which enables the user to give the robot verbal commands as well. Thus, their system supports the mechanisms of Joint Attention, cf. page 12 for details.

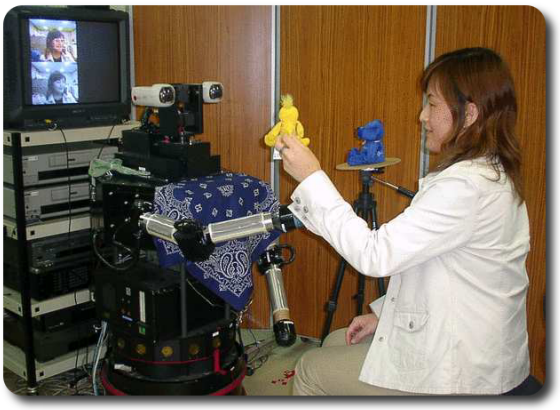


Figure 2.20: Robovie.
Image taken from [Nag04].

Concluding, although the presented system works quite fine for attentional shifts in combination with already known objects, the system is not able to learn a priori unknown objects. Last but not least, their approach does not offer a kind of Scene Memory which, however, is essential for autonomous tasks of the robot as stated above. This completes the overview of robotic platform that provide an implementation of an Object Attention. Next, the most important statements of this chapter are summarized.

2.4 Summary

In this chapter, different models of attention have been discussed in the beginning. In particular, models of Joint and Shared Attention as well as the idea of a Theory Of Mind have been introduced in the context of robotics. This overview has shown that the most sophisticated cognitive robots use an implementation of a Theory Of Mind to be able to cope with objects in a socially accepted interaction. This can be interpreted in a way that a Theory Of Mind supports the object-centered attention mechanisms best as the user and the robot are able to talk about the meaning of the referenced object. Hence, the robot might know what the interaction partner thinks about the object.

In the second part of this chapter it has been shown that different processing approaches using single as well as multiple input cues exist to support the Object Attention. To sum up, a broad agreement exists in the robotic community, that

only multimodal approaches are able to match the requirements of an intuitive and comfortable Human-Robot Interaction as it is more similar to scenarios humans are accustomed to. Here, especially Human-Human Interactions are usually the reference scenario.

At the end of the chapter, different robots have been presented which apply the attentional mechanisms described before. That section pointed out that only very few robots can actually deal with an Object Attention although most of them at least offer the necessary hardware prerequisites. This leads to the following final conclusion.

All approaches presented above lack the integration of a natural, intuitive, robust, and extensible object-centered Attention System. Some of them are good in vision processing, some are good in reasoning, and some are good in audio-based issues but a system that, ideally, copes with all three topics based on sophisticated approaches, does currently not exist. However, a lot of approaches can be used and integrated into the proposed Object Attention System in order to match the requirements for a natural Human-Robot Interaction.

The next chapter will, therefore, give an introduction on the techniques used for different input modalities as far as they are applied in the proposed Object Attention System.

3. Selected Modalities as Sources for Multimodal Object Attention

The previous chapter 2 has shown that only Object Attention mechanisms using multimodal input are suitable to face the challenging task of a natural Human-Robot Interaction. The given examples pointed out that Speech, Gestures, touch-sensitive user panels and easy to describe visual object properties (color, simple shapes) are the most promising candidates for robots with Object Awareness. However, not all of these features cover the preference for the integrated interfaces in the same manner as suggested in a preliminary evaluation presented by Khan et al. in [Kha98]. Their results of a questionnaire conducted with 134 participants about attitudes towards intelligent Service Robots showed that spoken language (82% preference) is nearly twice-preferred than written commands (45% preference). The study has also shown that touch screens are the second popular choice with 63% preference followed by gestures performed in front of the robot with 51% of preference. Therefore, a Personal Robot should provide a natural language-based interface for sentences, like, e.g., "Go to the sideboard in front of you", because this is more convenient for a human as an interface using a command language, like, e.g., "Forward 3 meter". That natural language is indeed of great interest for Human-Robot Interactions has been confirmed by a user study with 20 participants performed by J. Fritsch and colleagues [FWS05].

Besides these obvious interaction modalities, additional information about the user himself is useful in order to enable a pleasant atmosphere. For instance, an integrated face recognition could be used to identify the user and, thus, trigger the Object Attention System to complete an association for a personal object reminding event, e.g., "Axel, your black tea is waiting in the kitchen". This leads to the following structure for this chapter.

In accordance to the thesis overview presented on page 7, this chapter contains the details on the input modalities used for the proposed Object Attention System. In brief, these are all four modalities that are proposed by Khan et al. [Kha98]. Therefore, the chapter begins with modality information about the Person-Of-Interest (section 3.1), in particular, *Face Recognition, Leg Detection,*

Gesture Recognition, and *Speech Processing*. After the person-dependent input cues have been presented, the support for touch-sensitive panels (section 3.2) and written commands (section 3.3) is described. In particular, it is realized by the *Graphical User Interface (GUI)* that has been developed together with M. Saerbeck in his diploma thesis for the Object Attention System. Finally, the chapter closes with a brief summary (section 3.4).

3.1 Person Data

The processing of person-dependent data, like it has been proposed by S. Lang in [Lan05] and M. Kleinehagenbrock [Kle05] can be used, for instance, to increase the convenience level of a Human-Robot Interaction or to improve the accuracy of the determined Region-Of-Interest and the quality of learning objects, respectively. Thus, it is shown how face identification and the processing of distance values in combination with the evaluation of deictic gestures and speech are used by the Object Attention System in the context of person data. Although a great deal more processing like, gaze, haptic information, body heat, arousal of the communication partner, and so on, are conceivable, mostly, the selected features given above provide a reliable support in a human-like Human-Robot Interaction.

Identification and Distance of Communication Partners

Most people start a Human-Human Communication with a personal greeting while they address the communication partner with his name. In this way, often a more comfortable atmosphere is created. Furthermore, an identification enables the assignment of a specific object to an owner. Hence, if the user says something like "Bring me my red tea cup", the Object Attention System is able to reason about the information that Axel's red tea cup is addressed if the face identification has determined that Axel is the person who said the sentence. For a more robust identification the Object Attention System relies on two different methods. First, in case the user has said its name, the resulting symbolic speech entry is associated with the currently processed object instance. If the user does not state his name, a second method using a vision-based approach is evaluated as fallback solution. However, in the past it has been shown that neither the vision-based identification nor the speech-based identification is reliable enough to store the recognition results in the Long-Term Memory of the robot. Hence, to improve the reliability of the visual identification, a confidence value is considered by the Object Attention System for both approaches.

Experimental performance tests have shown that for the speech processing a reliability of 85% is realistic. This value corresponds to the word accuracy for the mobile robots used (cf. S. Hohenner [Hoh05]). The accuracy results from the Speech Localization by S. Hohenner [Hoh05], the Speech Recognition by G. Fink [Fin99], the Speech Understanding by S. Hüwel [HW06b], and the Dialog System by I. Toptsis [THH+05], and S. Li [LHW+05], respectively. For the vision-based face identification, evaluations have resulted in a value of 46% , according to T. Spexard [SHFS06]. The latter method of identification is provided by the *Person Tracking and Attention System (PTA)*, developed by S. Lang, M. Kleinehagenbrock, and T. Spexard [Lan05, Kle05, SHFS06] which includes a face

recognition and an identification submodule. This is discussed in more detailed, next.

For the recognition task of frontal faces as well as for faces that are up to 80° turned to the left or to the right, an approach proposed by Viola and Jones [VJ01] using simple rectangular regions as features is applied. Subsequently, a learning algorithm based on the *AdaBoost* approach introduced by Freund and Schapiere [FS95] which implements a chain of classifiers with increasing complexity is employed. As this only allows to recognize faces, but does not integrate an identification, the Person Tracking and Attention System uses the *Eigenface* computation proposed by Turk and Pentland [TP91]. In brief, this method represents for all faces a mixture of Gaussian functions with diagonal covariance in a particular Eigenface space. An example of such Eigenfaces is shown in Figure 3.1.

Figure 3.1: Eigenface representation for the user identification. Image taken from [Spe05].

Experimental results of S. Lang [Lan05] and M. Kleinehagenbrock [Kle05] have shown that besides face identification, the evaluation of distance values (Figure 3.2) helps to improve the quality of an interaction. For the Object Attention System this can be applied to analyze the user's actions as well. In order to access the measured values concerning the Person-of-Interest, the Object Attention System evaluates a continuously updated XML-document provided by the Person Tracking and Attention System. A fragment of such an XML-document is depicted below. In line 2 of the XML example the values are denoted by the tag "STATES" and its attribute "dst", respectively. The value of this attribute specifies the distance in millimeter which can be provided, e.g., by a laser range finder that is mounted on the robot. Besides, the value α [°] of the attribute "ang" with $\alpha \in$ [-90. . . +90] denotes the relative angle between the user and the robot.

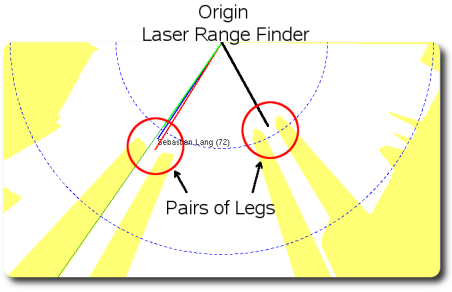


Figure 3.2: Laser-based distance scan of the Person Tracking and Attention module.

```
1  ...
2  <STATES ang="36.544791" dst="1818.802734" facing="NO" gazing="0" talking
       ="YES" walking="NO"/>
3  <IDENT id="3" name="Axel"/>
4  ...
```

The line 3 of the XML fragment illustrates that the Person Tracking and Attention System provides the tag "IDENT" which represents a consecutively numbered identification value and the name of the recognized user. If the user could not be identified, it is declared "unknown".

In the following, the gesture processing approaches that serve as additional input cues for the Object Attention System are introduced. Additionally, the upcoming

sections point out how the presented distance of the Person-of-Interest is applied to optimize the results of the gesture recognition.

Gesture Processing

Communicative and manipulative gestures are of great importance for the resolution of referenced objects occurring in an Human-Robot Interaction, cf. Strobel et al. [SIKM02]. Unfortunately, gestures can be carried out in many diverse ways. Therefore, the detection system needs to be flexible and adaptive in order to be able to recognize and track an adequate amount of gestures, according to Strobel et al. [SIKM02] and Guo et al. [GYJX99]. As it has been shown by J. Fritsch in [Fri03], visual scene context helps a lot to determine the performed action. However, as the approach by J. Fritsch is skin color-based only, it is vulnerable against varying lighting conditions. Furthermore, as the skin-colored regions are provided only by 2D-images, no depth information is available. But appropriate depth information about the current body pose can significantly improve the robustness of recognized actions. One method to provide a suitable context for the user's actions is to use a body model which allows to increase the quality of recognized gestures. Consequently, a 3D-Body Model is applied for the proposed Object Attention System which is described in the following.

3D-Body Model Tracking

The recognition of a gesture in a cluttered and natural environment is a highly complex task which consumes a great deal of computational power. Especially on mobile robots, where the computational resources are very limited it is desirable to reduce the amount of data to be processed. As a consequence, it is logical to reduce the search space by considering only postures, a human is able to perform. This can be realized by a corresponding Body Model. Furthermore, such a Body Model can help to disambiguate similar postures and, thus, can provide more accurate depth information. Hence, the 3D-Body Model developed by J. Schmidt [SKF06] is used to improve the performance of the subsequent gesture recognition.

As it is described by J. Schmidt et al. [SKF06], such a 3D-Body Model is quite suitable to support an Human-Robot Interaction although it relies only on a single monocular camera. However, the system used is not yet optimized and, thus, does not allow an online performance within the robot's architecture applied. Figure 3.3 illustrates the principle processing of the 3D-Body Model Tracking System while the following section schematically describes the processing during one exemplary iteration cycle.

In the beginning, after an image is captured ❶, it is preprocessed. This way, limbs, face, and hand regions can be detected. To do so, a skin color model in combination with edge and color cues is applied ❷. Furthermore, the edge filter helps to determine ridges in the given input image while this depend on the computation of the first and second partial derivative, respectively. In detail, the first partial derivative is used for the edge cue as it provides significant contrast changes only. Then, for the ridge cue, which depends on the image size of the limbs, additionally, a Gaussian image pyramid is created. This in turn is calculated by the distance between the regarded limb and the camera lens. Thus,

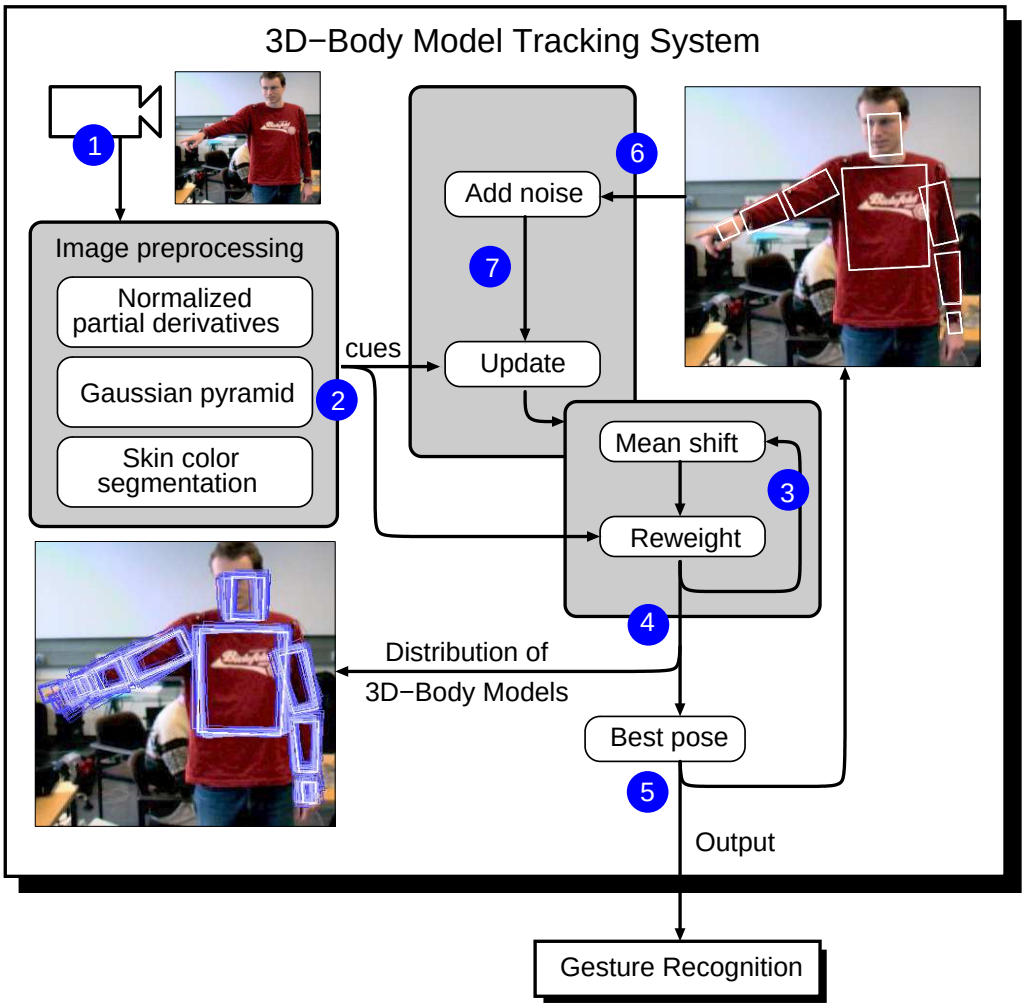


Figure 3.3: Schematic of the 3D-Body Model Tracking System.
See text for a detailed description. Image adapted from [SKF06].

single edge points can be suppressed while the image is analyzed for an existing limb. As a result of this preprocessing, a probability distribution is generated. This distribution is analyzed during multiple iterations of the *Mean Shift Algorithm* (cf. [SKF06]) ❸. Subsequently, the probability distribution is used to identify different modes representing diverse body poses ❹. As an outcome of the identification process, the Body Model with the highest rated body pose is selected ❺ and, additionally, used ❻ for following frames to provide a more reliable accuracy for new configurations of subsequent postures. Finally, a percentage of the generated particles (weighted sample points in space) is disturbed ❼ in order to allow an estimation of following body postures and movements, respectively.

The circumstance that the body tracker uses a single monocular camera only leads to a distance variance which depends on the gesture, environmental conditions (e.g., different lighting conditions), and observed movements of the user. Hence, a variance value of $\pm$ 200 mm is not uncommon. In order to partly compensate this variance and, consequently, get a more precise position of the limbs, a more accurate distance value provided by a 2D-laser range finder can be used, like it is mounted on the robot hardware that served for this thesis as evaluation platform. The equation 3.1 on the next page illustrates the correction term, where

$dist_{bt}$ represents the distance provided by the body tracker. $dist_{laser}$ is the measured value from the laser range finder after the Person Tracking and Attention System module has determined the Person-Of-Interest, and $dist_{corr}$ denotes the correction factor for the distance value. As an outcome, all relative distances between the limbs are scaled by the resulting distance correction.

$$dist_{corr} = \frac{dist_{bt}}{dist_{laser}} \tag{3.1}$$

To sum up, the presented 3D-Body Model approach matches the requirements for the subsequent gesture recognition process by generating hand trajectories that can directly be processed by the gesture recognition system. The Figure 3.4 illustrates an exemplary trajectory of a pointing gesture, like it has been produced during the evaluation of the Object Attention System.

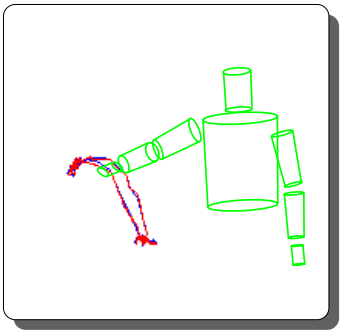


Figure 3.4: 3D-Body Model with an exemplary hand trajectory of a pointing gesture.

Consequently, the pre-processing of the 3D-Body Model framework offers two advantages. The first one is the support for real depth information which significantly improves the accuracy of all gesture-related coordinate values. The second advantage is that the gesture recognition system is less dependent on the error-prone skin color detection which is highly sensitive against varying lighting conditions and wooden surfaces in the robot's vicinity. In the next section it is described how the data provided by the body tracker is used for the trajectory-based gesture recognition.

Condensation-based Trajectory Recognition (CTR)

The gesture recognition module that, finally, provides the Object Attention System with gesture information is the *Conditional Density Propagation (CONDENSATION)-based trajectory recognition system (CTR)*, developed by N. Hofemann [HHFS05]. To do so, the CTR component supports two different modes, one for a 3D-based gesture representation and one for 2D-processing. The 3D-based approach applies the results, in particular, the trajectories of the previously described 3D-Body Model framework for the recognition of performed gestures, cf. Figure 3.4. Alternatively, the CTR module can directly deal with 2D-input images, see Figure 3.5.

In case of 2D-data processing, the CTR component utilizes a skin color-based region tracking in combination with a *Kalman filter* [May79] to extract the hand trajectories resulting from the movements of the user. In either case, 2D and 3D, after the trajectories are available, the further processing within the CTR module is basically the same. Briefly, the actual recognition is done by comparing the current motion with previously trained models and, thus, a specific action like a pointing gesture can be determined as soon as a certain threshold is exceeded. Nevertheless, the 3D-Body Model Tracking System provides a great deal more reliable data since it supports real depth information instead of empirically guessed values and, thus, the 3D-based gesture recognition leads to significantly better

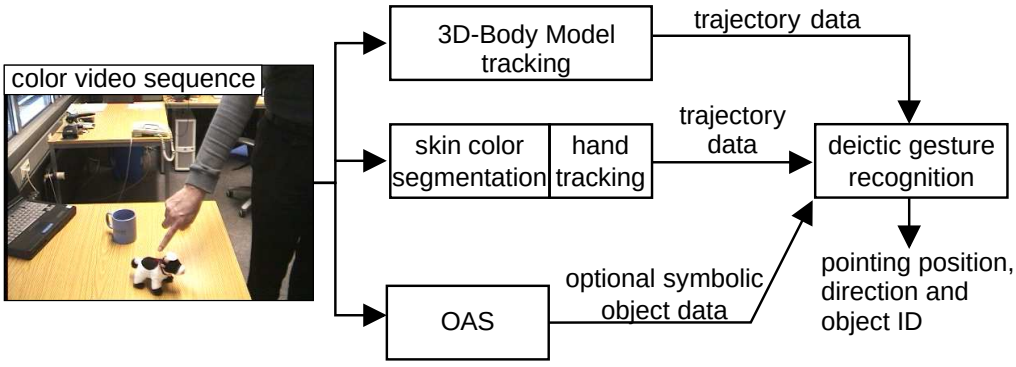


Figure 3.5: Modules for the deictic gesture recognition.
Image adapted from [HHFS05].

results. For the Object Attention System it does not matter whether the 2D or the 3D-based approach is used as it automatically detects the different input formats and dynamically initiates in each case the appropriate processing strategy.

The input that is provided by the CTR module contains the hand position or, in case of the 3D-approach, the hand, face, and head position. An excerpt of an XML document, like it is received by the Object Attention System is illustrated below. Here, the 3D-based variant is presented, as it contains more entries that are relevant for the Object Attention System than in the 2D-case. In the latter case, only the hand position, pointing direction, and gesture progress are considered.

```
1  ...
2  <TIMESTAMP>1156863546969</TIMESTAMP>
3  <ID>
4    <Origin Mod="GrabImg" Timestamp="">00236_1153819261_456.png</Origin>
5    <Origin Mod="BodyModelTracker" Timestamp=""/>
6    <Origin Mod="CtrXMLImport" Timestamp="1154591248898"/>
7  </ID>
8  ...
9  <RAW>
10   <STEP T="0">
11     <RIGHTHANDPOS X="2.418990" Y="-0.178610" Z="-0.258875"/>
12     <RIGHTHANDTIP X="2.245982" Y="-0.381262" Z="-0.299064"/>
13     <HEADPOS X="2.496483" Y="-0.054796" Z="0.204690"/>
14   </STEP>
15   <STEP T="-1">
16     ...
17   </STEP>
18   <STEP T="-2">
19     ...
20   </STEP>
21 </RAW>
22 ...
```

The XML example shows, solely, the data tags that are processed by the Object Attention System. Line 2 illustrates the time[1] just before the message is sent

[1] seconds since midnight UTC of January 1, 1970 (POSIX time), not counting leap seconds

by the CTR module. The three other timestamps (line 4 to 6) are necessary as the current implementation of the 3D-Body Model Tracking System does not allow an online performance. Hence, the "GrabImg" value denotes the consecutive number (236) of the captured image and the seconds (1153819261) with the corresponding milliseconds (456) in POSIX time. In line 5, the online timestamp of the 3D-Body Model tracker will be provided, as soon as it reaches real-time capabilities in combination with the CTR module. Until then, the line 6 provides the timestamp of the time when the XML data from the 3D-Body Tracking System has been recorded on hard disk.

As the recognition of the current posture is probability-based and, therefore, more or less disturbed, the last three time steps of the CTR are given, where the time between two subsequent steps is approximately 66,6 ms ($\widehat{=}$15 frames per second, like it is used by the CTR module). In order to minimize the noise influence, these three values are averaged by the Object Attention System.

Each time step consists of three tags which refer to the position within the 3D-Body Model as shown in Figure 3.6. Firstly, the position *RIGHTHANDTIP* (1) gives the coordinates of the fingertip. Secondly, the position *RIGHTHANDPOS* (2) which denotes the wrist of the right hand, and thirdly, the estimated position of the neck, denoted as *HEADPOS* (3). All values of the attributes *X*, *Y*, and *Z* (e.g., line 11) denote the distance between the user and capturing camera lens. In detail, the *Y* and *Z* values describe the deviation from the center of the camera image in the horizontal line, and the deviation from the center of the camera image in the vertical line, respectively.

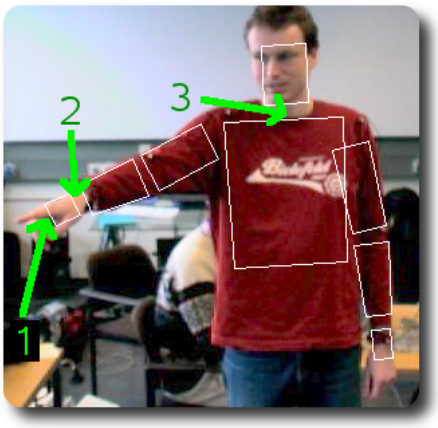


Figure 3.6: Illustration of the 3D-Body Model after the mean shift algorithm has been applied. Image adapted from [SKF06].

Besides the pure gesture recognition, the CTR module does have the ability to recognize actions as well, like drinking, making a phone call, typing on a keyboard, and so on. This action recognition, however, requires symbolic object information from the proposed Object Attention System that sends the object context to the CTR module, cf. [HFS04]. In detail, the object type (cup, bottle, keyboard, . . .), and the addressed object ID is sent to the CTR module while the information is sent as soon as the Object Attention System is informed by the dialog component about the currently discussed object.

The gained information about the Person-Of-Interest and his performed gestures is worthless as long as the point of time when an object reference should be resolved by the Object Attention System is unknown. Hence, the user's speech is used to resolve this timing dependency. The dialog system is able to provide the Object Attention System with the appropriate symbolic speech data which is presented next.

Speech

Symbolic speech and sound input plays an important role within the Object Attention System. It initiates the fusion process of gesture, speech, visual object appearance, and acoustic information of an object. But, before the symbolic data can be processed in the Object Attention System, the speech signal is preprocessed by the *Speech Recognition* of G. Fink [Fin99], the *Speech Understanding* of S. Hüwel [HW06b], and the *Dialog* module. For the latter one, two different implementations exist, which are both supported by the Object Attention System. The former slot-filling approach of I. Toptsis [THH+05] and the more recent implementation of S. Li [LHW+05] that allows an easier integration of context-dependent initiative models. Thus, it better supports studies concerning the issues, when the robot should say something. As the current developmental state of both dialog approaches offers the same functionality for the Object Attention System, only the more recent model of S. Li is regarded in the following.

When the communication partner is talking to the robot, the sound signal passes the Speech Recognition at first. It is based on a probabilistic *Hidden Markov Model* approach which generates raw symbolic speech data [Fin99]. Subsequently, this symbolic information is semantically interpreted by the Speech Understanding module [HW06b] which especially focuses on spontaneous speech as it is common in a Human-Robot Interaction. As an outcome, e.g., verbs or referenced object types can be determined and, thus, forwarded to the dialog component. The dialog module in turn decides on the basis of given utterances, whether the user announces that he wants to show the robot an object or otherwise the user wants to talk about an object that is already known by the robot. As a consequence, the dialog model currently supports two different orders that are evaluated by the Object Attention System. Firstly, an *Align View* command and, secondly, a *Focus Object* command.

The *Align View* command allows the robot's architecture to hand over the steering control of the camera to the Object Attention System. This enables the Object Attention System to lower the camera in order to get a better view on the user's hands. Additionally, the Object Attention System sends a response message to the dialog component. This is done, as soon as the camera is aligned and if the gesture recognition module is active, a confirming attribute "GestureExpected" is added to the corresponding XML document.

The order *Focus Object* causes the Object Attention System to align the camera as well if it is not already aligned. Additionally, the order provides the Object Attention System with information about the currently referenced object, as it is illustrated in the XML fragment on the next page. As the example shows in line 2, a timestamp is included which marks the time of the moment when the message was sent by the dialog component. Furthermore, an ID tag (line 4) indicates the current discourse ID within the dialog module. This ID is important for the Object Attention System in order to enable references to a specific query. In line 6, the name of the order is illustrated. If the user indicates a gesture, for instance, with an utterance like "This is...", the speech understanding and the dialog module expect a pointing gesture and, thus, the corresponding value in line 10 is set to 'yes'. Below line 10, the tag 'ObjectList' contains either none, one, or more than one objects that have been referenced, while each object is

denoted within a separate 'Object' tag, as shown in lines 12 to 15. Furthermore, the 'Object' tag contains the type of an object, like, e.g., bottle, cup, keyboard, and so on which is used for later object recognition tasks. Additionally, the 'Object' tag allows to embed various feature tags. The two most important ones consist of the type 'Color' which indicates the verbally specified object color and the type 'maybeGesture' providing the Object Attention System with the expected time of a performed deictic gesture. This timestamp refers to the moment when the user begins an utterance. As the speech understanding and dialog module assume that the user usually begins a sentence with "This is...", when he wants to refer to an object, it can be assumed that a corresponding pointing gesture is performed at that moment, as well.

```
1  ...
2  <TIMESTAMP>1130316022360</TIMESTAMP>
3  <ID>
4    <ORIGIN mod="DLG">4</ORIGIN>
5  </ID>
6  <NAME>FocusObj</NAME>
7  <STATE>ObjectAttention</STATE>
8  <DATA>
9    <OBJDESCR>
10   <GestureExpected val="yes" />
11   <ObjectList>
12     <Object type="cup">
13       <Feature type="Color" val="blue" />
14       <Feature type="maybeGesture" time="1130316022195" />
15     </Object>
16   </ObjectList>
17   </OBJDESCR>
18 </DATA>
19 </MSG>
```

To sum up, the presented XML representation of symbolic speech data allows an easy extension of the content by simple adding of further feature tags. Thus, the Object Attention System can be provided with various verbally given object information.

Besides the person-dependent information processing, like it is done by the Object Attention System, two more intuitive interfaces are realized in the proposed system. As the questionnaire results from the beginning of this chapter on page 33 pointed out are touch screen and textual input editor interfaces very popular for the instruction of a service robot, as well. For this reason, the GUI has been developed in cooperation with M. Saerbeck [Sae05] to provide such interfaces.

3.2 Touch Screen

A touch screen interface complements the input modalities considered by the Object Attention System, in particular, the gesture recognition module. Thereby, a touch screen might become necessary if the provided gesture data is not accurate enough or the gesture recognition fails at all. Such a failure can be caused by, e.g., varying lighting conditions or insufficient trained motion models which

are based on the hand trajectories. Another advantage of a touch screen interface lies in the possibility to give the user a direct visual feedback of the currently selected Region-Of-Interest. Thus, it becomes easy to recognize and, subsequently, to correct wrong selected areas. Even for multi-colored or partly occluded objects this is essential, as it is very difficult to automatically segment such objects correct. In general, that is the reason why most approaches select only a simple bounding box around objects instead of segmenting a fine-grained boundary around them as described in approaches presented by, e.g., Ghidary et al. [GNS+02], Kruijff et al. [KKBL06], or Wünstel et al. [WR06a].

To overcome these limitations, the GUI of the Object Attention System offers an interface for precise selections of referenced locations. This is enabled by displaying the current field of view of the object camera whereas the user can mark the center of the Region-Of-Interest with his finger, or instead with the mouse pointer.

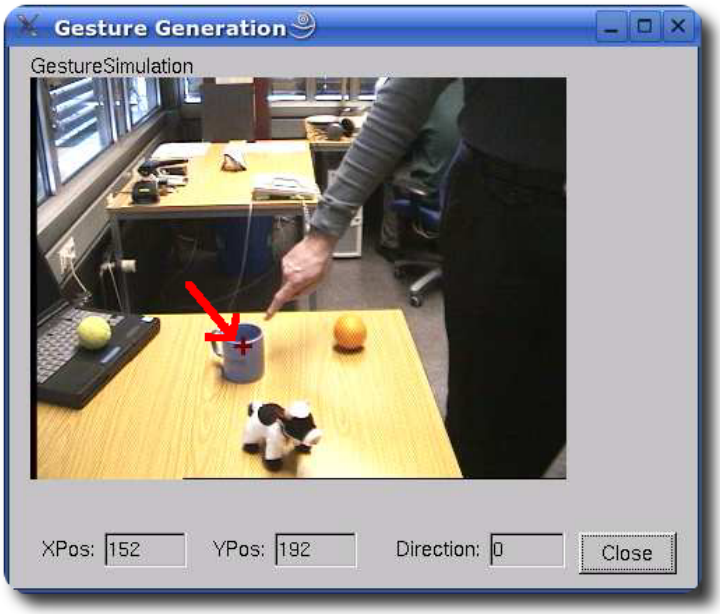


Figure 3.7: Touch screen functionality of the Object Attention System. Image adapted from [Sae05].

An example showing a manually selected Region-Of-Interest is given in Figure 3.7. The selected center of the Region-Of-Interest is marked by the red cross that is, additionally, highlighted by the red arrow pointing to it. For evaluation reasons, the user interface window, furthermore, contains information about the selected position in image coordinates. This is visualized at the bottom of the window. However, the shown text field for the pointing direction is currently not used.

To cover most of the preferred interfaces, the Object Attention System supports besides Speech, Gesture, and a touch screen interface, a fourth modality. In accordance to the most popular interaction interfaces presented by Khan et al. in [Kha98] (cf. page 33), the next section gives a brief introduction to the textual input editor of the Graphical User Interface.

3.3 Textual Input Editor

The integration of text fields usable for written commands offers three functionalities. Firstly, it enables the Object Attention System to be evaluated completely autonomous without a connection to other modules in a robotic system. This is of great advantage, in particular, for rapid prototyping. Secondly, the user is always

informed about the currently processed data and, thus, can more easily detect wrong recognized utterances due to the direct visual feedback. Thirdly, the text-editable areas support an interface for manually entered written commands. That way, a further modality is supported in order to make an interaction more comfortable for the user, for instance, if speech input is too difficult or the utterance is not recognized at all.

In its current implementation, the GUI supports two different modes for entering and querying object information. The first one is depicted in Figure 3.3 and shows the default view when an object is learned. At the top of the window, the user selects via different tabs which mode he wants to see or edit. On the left, the text fields for the symbolic speech data are present which offer a selection of the most important value, like color or object type. On the right, the slider "PixelAcceptThresh" can be used to refine the selected areas for the object view that was shown in the last paragraph. Thus, an insufficient learning result, e.g., due to varying lighting conditions can easily be corrected. Below the slider, a text field shows the currently processed unique object ID. At the bottom of the window, two checkboxes are given. Both are used to indicate whether the object information should be processed either from the Object Attention System or the GUI. The one called "From ShortMem" causes the system to process the data provided by other modules of the robotic system, for instance, the dialog component, which are stored in the Short-Term Memory of the Object Attention System. The second checkbox called "From ObjCommFlow" determines whether the mode of object learning is selected by the Object Attention System (checked) or by the GUI (manually entered by the user). Last but not least, the button "Go" causes the system to perform the computations in order to extract an object view. This button is, consequently, used if the manually entered values should be processed.

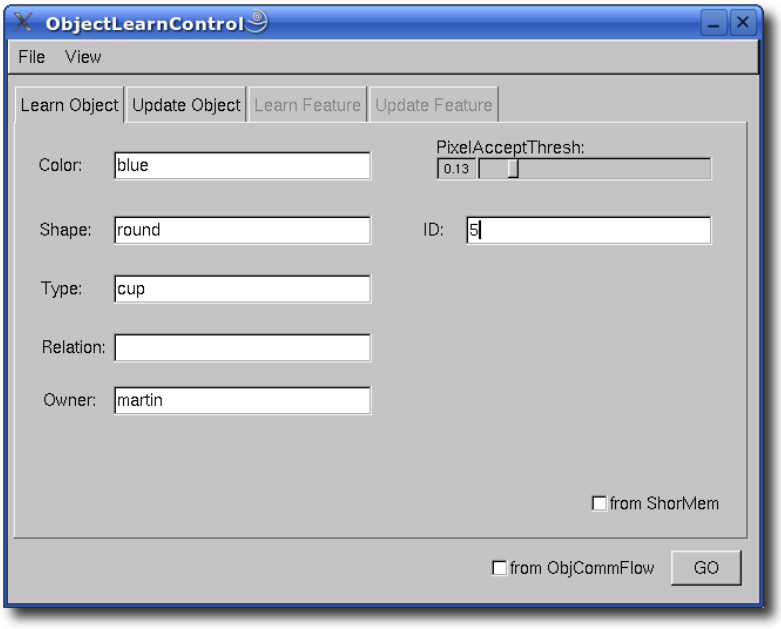


Figure 3.8: Graphical user interface used for written commands. See text for a detailed description. The image has been taken from [Sae05].

Concluding, the Object Attention System supports all four interaction modalities suggested by Khan et al. [Kha98]. After these aspects have been described, the following summary points out the main statements of this chapter.

3.4 Summary

In this section all user-related input modalities that are used by the Object Attention System were presented. At the beginning it was discussed, which are the most popular interaction interfaces for convenient Human-Robot Interactions. Then, the processing of user-dependent information (face identification, user position) was described in order to improve the accuracy of the Region-Of-Interest determination and to enable the connection between a user and an object, e.g., the extraction of the probable object's owner. Subsequently, the section about person-dependent information also included the gesture processing method used in order to recognize deictic pointing gestures. They enable the robot to pinpoint object references that are supported by gestures at the same time. After the discussion of gestures as input modality, the following section gave a brief introduction to the interface that allows to process symbolic speech data. In particular, the speech information that is supported by an interconnection of speech recognition, speech understanding, and a dialog component. In this way, the Object Attention System is provided with important object information verbally given by the user, like, e.g., the object's color. Thus, further processing within the Object Attention System and, hence, the extraction of an object view becomes possible. As last input modality, the graphical user interface including support for a touch screen and written command has been presented which completes the discussion on the different modality interfaces realized.

In the following chapter, the internal processing mechanisms of the proposed Object Attention System is presented. In particular, the chapter describes how the object information is actually fused, rated, and evaluated.

4. Development of an Object Attention System

This chapter describes the main contribution of this thesis. Therefore, the communication, data representation, and processing strategies are presented. During the last chapters, the principles of Attention for a successful Human-Robot Interaction has been described as well as a selection of useful modalities that support a convenient interaction. It has been pointed out that a scene analysis that enables a robot to robustly learn single objects demands for multiple modalities (e.g., Speech, Gestures, . . .). These modalities in turn need to be joint with each other in a spatio-temporal sense. However, neither the explicit object segmentation of a scene nor the temporal analysis has been discussed so far and is, thus, part of this chapter. Due to its specific character, the learning of objects and the spatio-temporal dependencies are presented in a separate related work section. This leads to the following outline for this chapter.

In the following, an overview of the related work for object learning is described (section 4.1) together with a brief introduction to the robot platforms used (section 4.2). In compliance with the thesis overview (Figure 1.3 on page 7), then, the realization of a unified interface follows (section 4.3) that is directly connected to the implemented Short-Term Memory (section 4.4). It is responsible for accurate data representations of visual and auditory object information (section 4.5). Within the Short-Term Memory the proposed Object Attention System decides whether the object referenced by the user is unknown to the robot or if it is an already known object. The underlying control mechanisms has been realized by a Finite State Machine for the integration of these two different processing strategies (section 4.6). Finally, a concluding short summary is given (section 4.7).

4.1 Related Work

In this section, a presentation of the most relevant related work in the context of object learning for robots follows. Therefore, the main advantages and disadvantages of the different proposed systems in relation to the implementation of the

proposed Object Attention System are discussed. At the end of this section, a short introduction to field of temporal dependencies between different modalities, e.g., gesture and speech, are given.

Approaches for Online Object Learning

For a natural Human-Robot Interaction and, consequently, the development of the Object Attention System, the aspect of object learning is one of the major issues as described by, e.g., Nicolescu, Mataric, Taylor, and Kleemann [NM01, TK04, SGW+06]. For instance, like Steil and Wersing [SW06] summarize the trends for online learning in cognitive robotics, hard-wired behavior sequences in a cognitive robotic architecture limit the capabilities of flexible object learning. They point out that especially biologically-inspired cognitive vision approaches are the most promising ones for online learning tasks between a teacher and a robot. They conclude that due to the need for a highly reactive and adaptive behavior of the robot in a natural Human-Robot Interaction, traditional approaches using, e.g., *Multi-Layer Perceptrons (MLP)* or *Support Vector Machines (SVM)* fail due to their online performance, cf. Kirstein and colleagues [KWK05, WKG+06].

Kirstein et al. [KWK05] therefore follow the paradigm of a separation into a Short-Term Memory for fast reactive learning tasks and a *Long-Term Memory* for storing persistent but not time-critical information. This separation is useful for the Object Attention System as well, as the object learning takes place in a few seconds during a Human-Robot Interaction while for autonomous interactions (e.g., fetch-and-carry tasks) by the robot, the object information learned can be transferred to the Long-Term Memory. For the realization of the memories, Kirstein and colleagues use a slightly modified supervised *Learning Vector Quantization (LVQ)* algorithm. This enables their system to create a Short-Term Memory using similarities with an adaptive collection of object view templates. Subsequently, the learned feature representation is used for an incremental *LVQ* in order to accumulate the features into the Long-Term Memory. Thus, it is possible to continuously train the Long-Term Memory which usually results in a decreasing classification error. They have demonstrated the feasibility of their approach by learning 50 objects in about three hours while the remaining classification error was about 6% using color and shape and an 8% error by using shape features only. These results are very promising, although Kirstein's approach does not match the requirements of a natural interaction as no commonly used interfaces, like a gesture or speech recognition have been integrated.

In [BSD03, BSZD06], Becher et al. propose an interactive Object Modeling System for semi-autonomous learning of object models. In this context, the practicability for these object models is discussed with regard to a convenient Human-Robot Interaction as well. For instance, such object models have features, like "is transportable" or "is a container", and the models are grouped into classes. Consequently, each single model represents an instance of a certain class, like "cups". The features contain several attributes, for instance, the fill state for containers with corresponding attribute values, like the numeric value for its fill state (e.g., 10%). In order to enter this information they use a graphical user interface that allows to enter the names of attributes as well as their values. For the sensory input they use a 3D-laser scanner and a high-resolution color stereo camera

system which enables a visual analysis of an exposed object within the restricted modeling area as shown in Figure 4.1.

Figure 4.1: Object modeling area of Becher et al. [BSZD06].

In order to enable further processing of the manually entered object information, the attributes and their values are represented as *Extensible Markup Language* [Wor06a] encoded text. This object text is inserted in a database for modeling the world knowledge of the system. Summarizing, the proposed system for object modeling is able to generate detailed object models that can be used for later recognition tasks. However, the system is only partial usable for a convenient Human-Robot Interaction as its current implementation does not support naturally spoken language or gestures without additional sensors.

An interesting approach that does not rely on such a restricted modeling area is presented by Hois and colleagues [HWBR06]. As their approach for object-related modeling does support speech processing, it is more suitable for the requirements of a natural Human-Robot Interaction. As sensor they use a 2D-laser scanner mounted on a pan-tilt unit in order to recognize objects in a domestic environment.

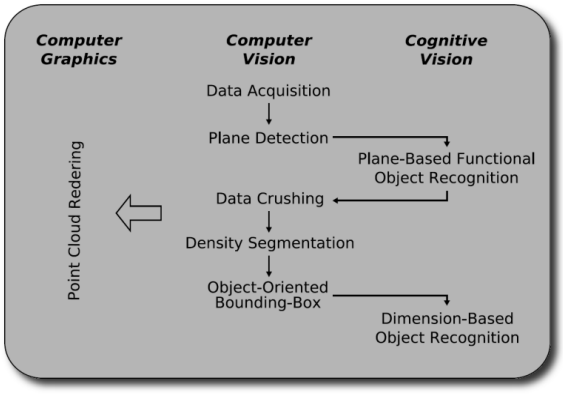
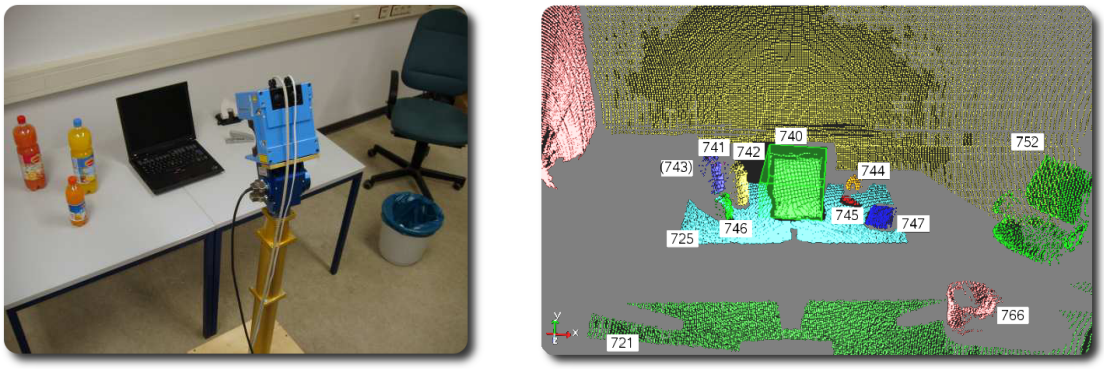


Figure 4.2: Object Recognition using Cognitive Computing system. The figure has been taken from [HWBR06].

Their proposed recognition system *ORCC (Object Recognition using Cognitive Computing)* depicted in Figure 4.2 is able to segment a cluttered scene based on a visual 3D-laser scan by putting a bounding box around single objects. The processing starts with a *Data Acquisition* step performed by their 2D-laser range finder which is tilted in order to achieve a 3D-laser scan. The resulting depth image provides an accuracy of approximately 1 cm. Then, in a subsequent *Singular Value Decomposition (SVD)* step [WR06b] which is combined with a region growing algorithm, planes in the scene are extracted. After this first *Plane Detection* for the complete scene content, a *Plane-Based Functional Object Recognition* is initiated in order to gain a coarse separation of objects from their environment, like walls or table surfaces. Therefore, the object models offer attributes, like *orientation* or *object size*. Furthermore, these attributes describe relations between the different models, for instance, *distance* or the *deviation of the orientation from the horizontal plane*. For the following object detection, all points within objects are removed in a *Data Crushing* step as structural information would not allow a sufficient object segmentation. For the segmentation

Hois et al. then use a *Density Segmentation* approach to enable the detection of already known shapes, but also of unknown objects. However, the previously gained knowledge about the environment, like tabletops is also used to improve the results by restriction of the scenario. In a last processing step, an *Object-Oriented Bounding Box* is set around each segmented object. Hence, a *Dimension-based Object Recognition* can be initialized using these bounding boxes.

An exemplary processing result is shown in Figure 4.3. Image 4.3(a) illustrates the scenario. In the foreground, the laser range finder mounted on a pan-tilt unit can be seen. In the right image 4.3(b), the final 3D-representation is depicted. It shows the already separated and labeled objects together with their bounding boxes. Next, a brief Human-Robot Interaction scenario is described that illustrates how this 3D-representation is applied for the mobile robot used.



(a) 2D-scene. (b) 3D-scene.

Figure 4.3: 2D- and 3D-representation of a scanned scene.
The images have been adapted from [HWBR06].

As their purely vision-based approach is often not sufficient to let a mobile robot interact or navigate in an unknown environment (cf. section 2.2.2 on page 20), the *ORCC* system of Hois and colleagues combines the visual information with naturally spoken language. Thus, their system allows to correct wrong learned object classifications and instances. Furthermore, it can deal with spatial relations. These features are supported by a linguistic component, which is divided into two phases, the *training phase* and the *action phase*. During the training phase, the user is able to associate object type labels that could not be determined automatically. These labels are added to a domain ontology acting as knowledge base for the scene. In combination with the depth information from the laser range finder, the system can be asked about spatial relations between objects. First evaluation results for object recognition and the determination of spatial relations are presented in [WR06a, HWBR06]. These show that they have developed a feasible approach of a cognitive vision system connected to an ontology-based representation with linguistic support.

Summarizing, the approach of Hois and colleagues is quite sophisticated but it is not yet suitable for a natural Human-Robot Interaction because of the following reasons. The 3D-scan of the scene is very time-consuming. Thus, this model is more suitable for non time-critical offline object learning than for a dynamic

Human-Robot Interaction scenario. Furthermore, the user interface is rather limited as it currently includes speech only and even more their system is not able to cope with additional object attributes, like color. As depth is difficult to verbalize for humans (e.g., objects that are positioned in a 45° angle to each other, are they placed behind, side by side or both?), a gesture recognition can help to resolve such ambiguities. Last but not least, depth is currently the only input. For aspects, like partial occlusions or objects standing directly next to each other, the scene analysis will fail for these objects. In this context other features, for instance, shape or color could significantly improve the recognition results of the system.

An approach that overcomes some of these constraints by using visual object information as well, but also speech and gesture data is presented by F. Lömker in [Löm04]. He utilizes a static scene in order to visually learn objects. The visual learning algorithm is based on comparisons of histograms. This becomes available by the use of difference images that were calculated as soon as the user picks up an object and, thus, the system can determine which image part represents the object. In this way, the approach produces reliable output, however, it does not consider spatial issues. Hence, a second method based on graph-matching, like described by, e.g., C. Bauckhage and colleagues [BBS04] has been implemented by F. Lömker as well to overcome these limitations. As the evaluation of the system shows, the approach of F. Lömker matches many of the requirements for a cognitive motivated object learning architecture. However, it does neither consider spatial object relations nor object features, like *size* or *depth*. Unfortunately, a domestic domain often requires the analysis of these features, e.g., due to its cluttered character. Additional information is often the only way to resolve ambiguities. Approaches that are more focusing on the processing of context information are therefore discussed in the next paragraphs.

As it has been shown (e.g., Dickinson [Dic99], Triesch and Eckes [TE05], and E. Braun [Bra06]), a large amount of object recognition approaches already exist that can eventually be adapted for the use in robots.

A survey for inferencing generic object models based on examples is, e.g., described by Keselman and Dickinson in [KD05]. Such generic models are especially useful for autonomous interaction tasks of a robot, as they eventually allow an improved recognition of the objects that have been learned during a former interaction with a user. Thus, it makes sense to discuss the approach of Keselman and Dickinson in detail. They point out that during the 90's, appearance-based modeling became more and more popular in order to deal with complex, scaled, rotated, occluded or translated objects. One of the most challenging tasks regarding the recognition of objects is the still existing representational gap. This gap occurs as it is usually assumed that a 1:1 correspondence between the image and its model exists. In particular, they state that a saliency in the image does not automatically imply that it is contained in the model as well. Thus, the model needs to be represented as abstract as possible and, therefore, be able to abstract from salient details anyway. They propose an approach to develop the so-called *lowest common abstraction (LCA)* image, whose principles can be explained on the basis of Figure 4.4. As the illustration shows, the construction of an LCA image underlies a hierarchical concept. First, Keselman and Dickinson segment the real images, as shown in the middle row of Figure 4.4.

This enables a formation of a region adjacency graph γ , while it is assumed that every image is oversegmented. Then, a lattice λ is formed by merging all possible sequences of region adjacency graphs. Each image forms its own lattice where the bottom is the original region adjacency graph and the top τ of the lattice is the silhouette of the object. The latter one emerges from merging all regions into a single region. Now, a *common abstraction* is defined in such a way that for any two nodes of two different image lattices (λ_1, λ_2) their corresponding graphs (γ_1, γ_2) are isomorphic. Hence, each τ_n of every image is a common abstraction. The LCA (top row of Figure 4.4) is then defined as a common abstraction whose graph has got a maximal amount of nodes. In order to improve the performance of the search for an LCA, they search only for intersections of two lattices $\lambda_{1,2} \in \lambda_n$.

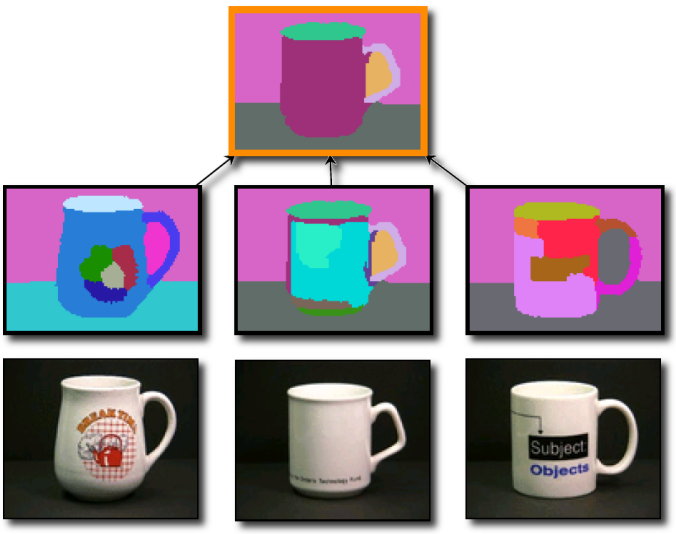


Figure 4.4: Lowest common abstraction image. Image has been taken from [SKD+06].

The benefit of the calculation of an LCA image is the creation of a generic object model that can be mapped on new scene images in order to recognize already learned objects. Nevertheless, object recognition with image-based appearance models still proves as prone to error. As a consequence, more and more algorithms are based on interest feature points, like the SIFT features by Lowe [Low04], or shape matching algorithms [BBM05] introduced by Berg and colleagues. Especially the SIFT features by Lowe [Low04] became an excellent approach to recognize interesting feature points in images. Thus, they are applied in the proposed Object Attention System as well.

One last aspect that has been considered for the implementation of the multimodal Object Attention System copes with the temporal relationships between the modalities used. In this field of research a lot of experiments have been conducted. The following section gives a brief overview of the most important issues as far as they concern the Object Attention System.

Temporal Dependencies between different Modalities

Deictic gestures often accompany a verbally specified object reference during a Human-Human Interaction. This behavior has successfully been observed or at least been evaluated during Human-Robot Interactions as well, as user studies have shown, e.g., [Kha98, FWS05]. The temporally correct relation assignment of different modalities (speech, gesture, object sound, and visual object appearance) is, therefore, one key feature of the proposed Object Attention System. Consequently, a lot of effort has been spent on the implementation and the analysis of temporal connections. For this reason, the Short-Term Memory is responsi-

ble for the temporally correct data fusion. To summarize the modalities that need to be aligned, all considered features are listed below:

- (Deictic) speech
- Deictic gesture
- Distance of the user to the robot
- Auditory object appearance
- Visual object appearance (Position)

In the following, these issues are discussed in detail in the following subsections.

Correlation between Gesture and Speech

The accurate timing of a deictic pointing gesture related to a specific utterance is the linchpin for the determination of the correct Region-Of-Interest. The relevance of object references that are supported by gestures, thus, has been a research topic for several years, as described by A. Kranstedt and colleagues [KLP+06], S. Wachsmuth [Wac01], or Kendon [Ken04]. Concluding, the literature shows that diverse time-dependent thresholds have to be considered in order to get the best accuracy for Region-Of-Interest. As a consequence, it is useful to orient by experiments that investigate the temporal dependencies between different modalities. Especially, gestures and speech should be related and synchronized to each other as they are the most volatile features in comparison to visual object appearances or written commands. In the following, experiments that are related to this topic and that have been conducted in a robotic setup are discussed.

The experiments that are described next, concern the multimodal fusion of 3D-pointing gestures and speech and have been conducted by Holzapfel and colleagues [HNS04] who present user studies with 7 participants. In their scenario they used a kitchen environment that allows to interact with the humanoid robot ARMAR by using speech and gestures. As this scenario is part of the Home Tour Scenario, like it has been described in this thesis, the results are considered to be adaptable to other robots as well. Consequently, the resulting time value relations have been used for the Object Attention System. An overview of the measured values is given in Figure 4.5 on the next page.

The results of the conducted experiment are based on 89 utterances accompanied by simultaneously performed gestures as it is illustrated in the upper part of Figure 4.5. The upper diagram describes the temporal correlation between the manually annotated begin of a deictic gesture and the begin of an utterance. As the diagram shows, most gestures began between 0.52 seconds before and 0.7 seconds after the speech began. Subsequently, Holzapfel and colleagues investigated the correlation between a deictic gesture and a deictic word, like "this" or "that". The corresponding results of this investigation are shown in the lower diagram of Figure 4.5.

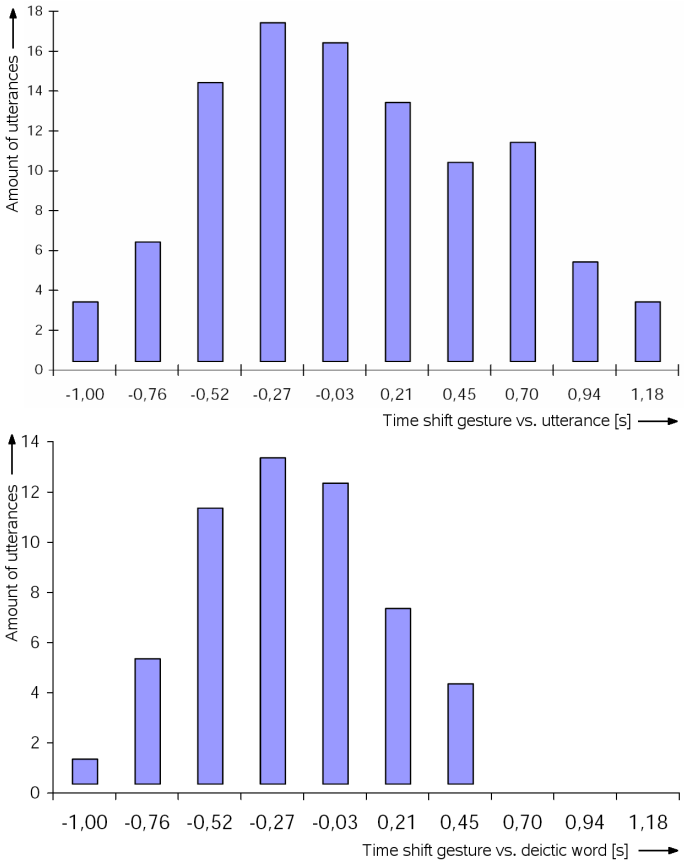


Figure 4.5: Time-correlation between deictic gesture, speech, and deictic speech. These figures have been adapted from [HNS04].

Here, a subset of the evaluation data has been used that contains 53 utterances. It can be seen that a close relation between a deictic word and a spoken utterance exists which is also verified by experiments of Sugiyama et al. [SKI+05] or McNeill in [McN92]. The data of Holzapfel, however, has been estimated with a normal distribution which results in a mean value of -0.3 seconds and a variance of 0.14 seconds, as it is described in [HNS04]. This indicates that the user often points to an object 0.3 seconds before he speaks his deictic word. However, these experiments concern only the correlation of gesture and speech. The duration of the gesture during its hold phase is, nevertheless, very important as well. Especially for the gesture recognition it is essential to detect the end position of a gesture as otherwise all subsequent calculations, performed by the Object Attention System for the determination of the Region-Of-Interest are getting inaccurate. Holzapfel investigated the aspect of the ending gestures in his experiments, too (Figure 4.6).

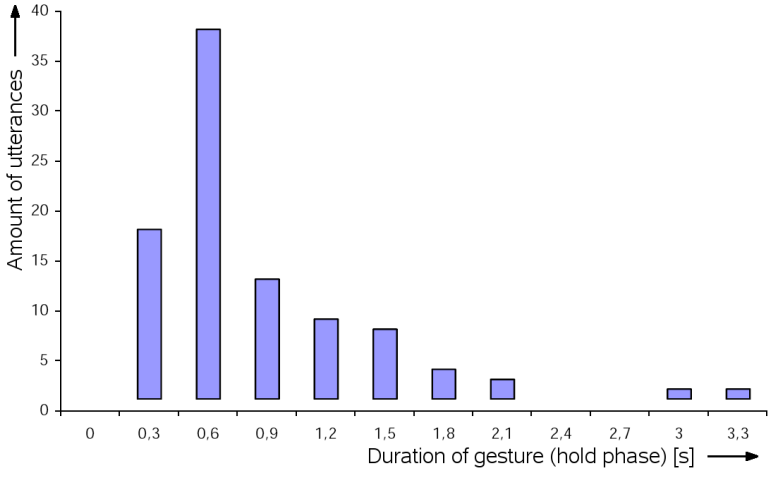


Figure 4.6: Duration of a deictic gesture during hold phase. The figure has been adapted from [HNS04].

The diagram illustrates that most of the gestures lasted no longer than 1 second while a significant peak in his evaluation shows that nearly half of the gestures investigated, were performed with an approximate hold phase of 0.5 seconds. But, the histogram also shows that every fifth gesture had a hold phase of approximately 0.3 seconds. However, further experi-

ments with virtual characters, conducted by S. Kopp [Kop03] and T. Sowa [Sow05] have shown that this time-dependent relation does not necessarily always apply for a Human-Machine Interaction. This proves that a high variability for deictic gestures exist, influenced by the task and the user.

Besides the relation between speech and gesture, it is useful to consider the presence of the user and his position as well (cf. page 38). In this way, the accuracy of the calculations for the Region-Of-Interest within the Object Attention System can be improved. Therefore, this issue is discussed next. However, since the proposed Object Attention System is an innovative approach with regard to the diversity of different modalities processed for Object Attention, no comparable related work is available yet. Nevertheless, as these time-related issues are semantically connected to the temporal dependency between gesture and speech, they are briefly introduced in this related work section as well. The same applies to the following paragraphs describing the relation to object sounds and object views. A more detailed description, however, is given later on in the subsequent sections of this chapter.

Location of the interaction partner

The position of the user at a particular time results from the Person-Of-Interest that is calculated by the Person Tracking and Attention module [Lan05, Kle05, SHFS06]. In particular, the distance of the Person-Of-Interest in relation to the robot is measured by a laser range finder, cf. page 35. Due to its measurement frequency of approximately 5 Hz (200 ms), the corresponding relationship between this distance and the estimated timestamp of the utterance can not be determined more accurate than these 200 ms. However, this has been proven to be accurate enough for a clear assignment (User Position $\Leftrightarrow$ Utterance) although a time span of $\pm$500 ms would be sufficient, too, as normally an interacting person does not move that much in half a second. Besides these temporal aspects for the Person-Of-Interest, the referenced object itself can provide useful information as well, like its sound. Therefore, a short description on this detail is given next.

Object sound

Capturing the correct moment in time in order to extract an object's sound is very difficult, as objects usually do not indicate the beginning and the end of the sound producing period. Of course, exceptions exist, like, e.g., alarm clocks that often additionally illuminate their display while they are beeping but that is barely a reliable aspect. Fortunately, the user can help with his utterance and his gesture to let the robot capture the correct time span as well as the direction from where the sound is generated. To match these requirements, the Sound Collector has been developed for the Object Attention System which establishes a temporal assignment between the user's utterance and the object's sound.

This Utterance $\Leftrightarrow$ object sound relation in a temporal sense has been chosen as speech offers the possibility to bind the begin and end time of a sound recording to specific key words or sentences, like "Listen to the following object sound" and "Now stop recording the object sound". Besides the audio-based object analysis does the visual object appearance provide valuable features, too.

Object view

As the object camera continuously captures images, it has to be considered which image needs to be extracted from the continuous video stream in order to get a clear object view. In this context, the Object Attention System offers two possibilities. On the one hand, the user can select the Region-Of-Interest just by using the touch screen interface, which has been described in section 3.2 on page 42. Although this provides the advantage of a direct feedback for the user, a pointing gesture offers an additional input cue and, thus, supports a more convenient Human-Robot Interaction. Therefore, the gesture is evaluated and, consequently, the camera can be aligned on the referenced location. However, it has to be questions how the system can determine whether the gesture is complete. For the Object Attention System this is solved by a continuous validation of the current camera position. As soon as the camera stops its motion it can be assumed that the camera has reached its destination position. Nevertheless, the evaluation of the timestamp when an image has been captured can increase the robustness for an accurate temporal assignment. In particular, this image timestamp could be put in relation to the gesture timestamp. However, due to technical constraints this has not been realized so far.

Concluding, it can be said that the different input information arrive at different time which demands for a buffering of the data. This is realized by the Short-Term Memory within the Object Attention System. But before the implementation details of the proposed Object Attention System are described, the robotic hardware platforms used are presented next with a focus on the sensors that provide data for the Object Attention System.

4.2 Hardware Platforms used for the Integration of the Object Attention System

The use of real robots for the development of the Object Attention System is essential mostly for two reasons. First, it enables a more robust implementation as it allows to evaluate the proper functionality of the Object Attention System. Secondly, the use of real robots demonstrates that the system not only operates in a laboratory environment, but actually on the intended field of application. While in the beginning of the development phase of the Object Attention System only one mobile platform existed, later on a stationary anthropomorphic robot became available as well. As a consequence, the proposed Object Attention System has been enhanced for the use with this robotic platform as well as it partly offers different sensors. Thus, the Object Attention System can be used with human-like sensors only and does not necessarily rely on artificial sensors, like the laser range finder any longer. The following paragraphs, therefore, briefly describe the different hardware used.

The Mobile Robot BIRON

One of the first steps during the development of the Object Attention System consisted in the examination of the existing robot platform with regard to the needs for the integration of the Object Attention System in the robot. It turned out

that for several reasons, e.g., insufficient computational power, it did not match the requirements for a successful integration. Thus, an enhanced new platform has been assembled. As a consequence, not only the Object Attention System but also further modules, like the gesture recognition by N. Hofemann [HHFS05] could be integrated as well. Image 4.7 shows the new mobile *Bielefeld Robot Companion (BIRON)*. It is based on a *Pioneer PeopleBot* platform from the company *ActivMedia*. Next, this hardware platform is described in more detail.

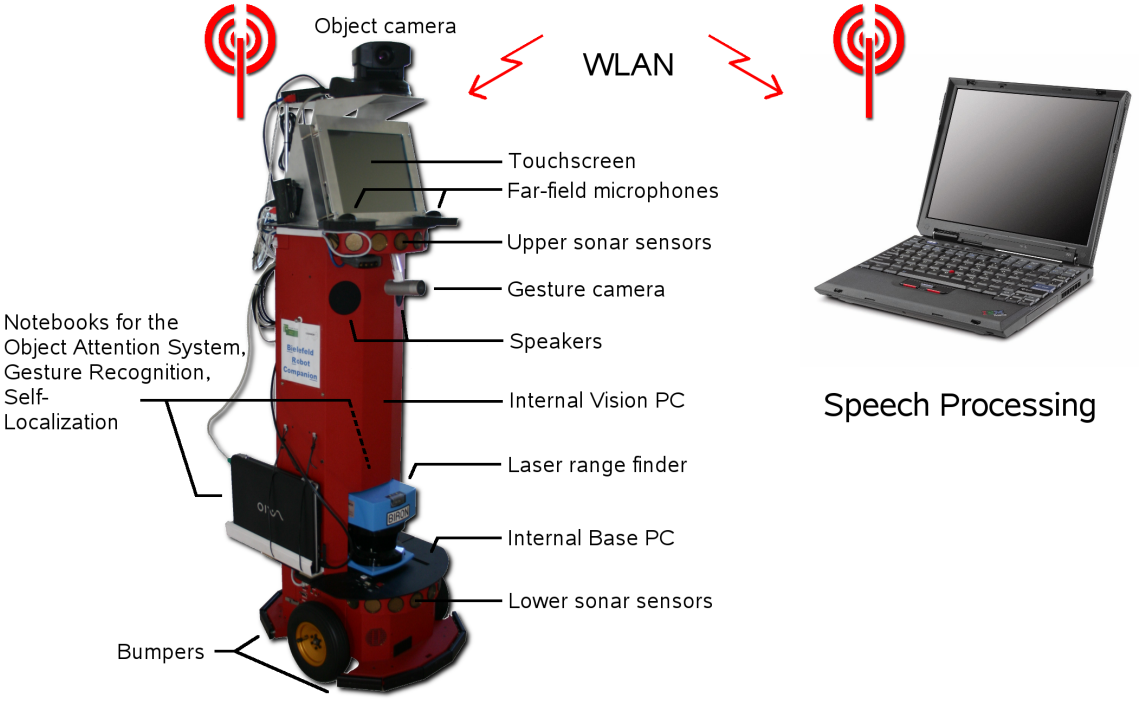


Figure 4.7: The Bielefeld Robot Companion (BIRON) of the applied computer science group at Bielefeld University. The wirelessly connected notebook depicted on the right is mainly used for speech processing tasks, e.g., the dialog system.

On its top, a Sony Evi-D31 pan-tilt camera is mounted at the front side at a height of approximately 1,42 m. It is the most important visual sensor for the Object Attention System as it is used to identify objects and, therefore, it is described more detailed than the other sensors. Besides, it is also used to identify the face of the current interaction partner (cf. 3.1 on page 34). The camera supports a resolution up to 768×576 pixel (PAL) and is steerable for about 100° to the left and as well as to the right, and for about 25° in each vertical direction. Its lens provides a field of view of up to 37,6° × 48,8° in its widest zoom position while the maximal ×12 optical zoom provides a perceptional field of view of 3,2° × 4,3°. The zoom functionality is approximately linearly adjustable over an interval divided in approximately 1000 steps. The camera is attached to a mounting that includes a display with a touch screen. This display is used for maintenance tasks as well as for visual feedback for the user during a Human-Robot Interaction. At the left and right bottom of the display two far-field microphones are mounted which capture the environmental sound for the speaker localization as described by S. Hohenner in [Hoh05]. Additionally, in a calm environment these microphones can be used

to capture the user's speech which makes the use of a close-talking microphone headset dispensable.

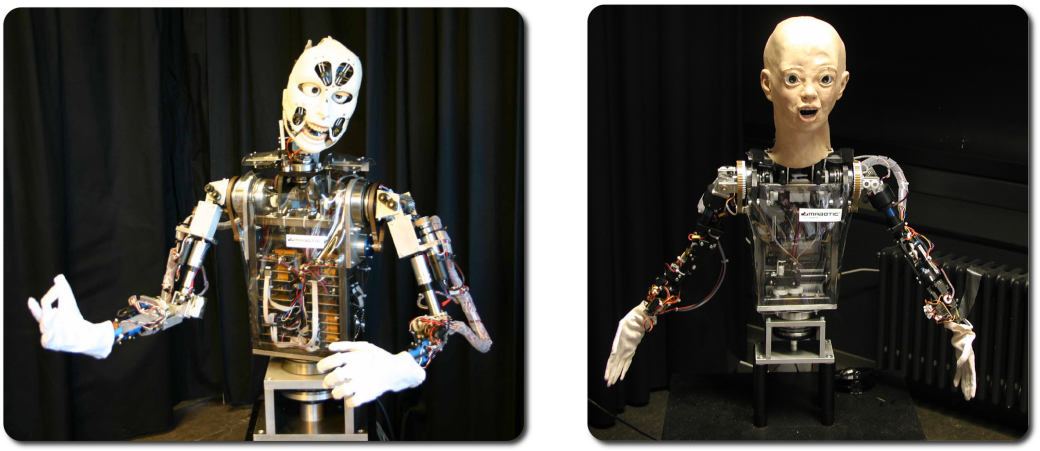
Beneath the display, a sonar ring including 8 sensors is mounted while an equally equipped second sonar ring is mounted at the base of the robot. These sensors can be used for, e.g., collision avoidance, but as they produce a clicking sound in audible frequency they are disabled as they would influence the speech recognition results. Directly under the upper sonar ring, a monocular Apple iSight camera is mounted which is used for the body tracker and the gesture recognition, cf. section 3.1 on page 36 and the following ones. Alternatively, a stereo camera can be mounted in order to get more accurate depth values. Also embedded into the tower casing of BIRON are two speakers that are used for its utterances during an interaction phase. Besides, the tower also contains an industrial computer that is mainly used for the visualization on the display and the Person Tracking and Attention System by M. Kleinehagenbrock [Kle05], S. Lang [Lan05], and T. Spexard [Spe05, SHFS06]. At the bottom of the tower casing two mountings are attached, one on each side in order to equip the robot with two notebooks that perform the computations of the Object Attention System, the gesture recognition, and a self-localization which is currently under development. Between the two laptops, the blue-colored laser range finder is shown. This 2D-laserscanner measures distances of the environment within a 180° plane scan at an approximate height of 30 cm. Thus, pairs of legs can be detected by the Person Tracking and Attention System by using an appropriate heuristic as described by S. Lang and J. Fritsch in [Lan05, FKL+03].

The basis of BIRON holds a motor which enables BIRON to move forward and to rotate. Since no sensors are oriented to the back, a backwards movement is disabled and so are the black bumper switches at the very bottom that can, for instance, be used for emergency stops if BIRON collides with an obstacle. Within the basis, the power supply consisting of high-capacity batteries and a second industrial computer are contained. This computer is connected to the two microphones beneath the display and to the motor controller board. Additionally, it serves as host computer for the wired network on the robot and the wireless network that is used to connect one or more notebooks that are used for, e.g., the speech processing as shown in Figure 4.7.

The technically oriented platform BIRON is due to its sensors not ideally designed to simulate the behavior of humans and, therefore, restricts the fields of application of the Object Attention System. Thus, the Object Attention System has been extended for the use with anthropomorphic robot platforms as well.

The Anthropomorphic Robot BARTHOC

In this section, the anthropomorphic robot *Bielefeld Anthropomorphic Robot for Human-oriented Communication (BARTHOC)* [HSF+05, SHFS06] is described. It has been developed by the company *mabotic* and is mainly used for user studies with human-like robots. The applied computer science group of Bielefeld University has got two exemplars of BARTHOC, an adult-like one (*BARTHOC senior*) shown in Figure 4.8(a) and a second one with child-like dimensions (*BARTHOC junior*), see Figure 4.8(b). The latter one is shown with a skin-like mask for the face while an appropriate mask for BARTHOC senior exists as well.



(a) Full body view of BARTHOC senior. (b) BARTHOC junior with skin mask.

Figure 4.8: The Bielefeld Anthropomorphic Robot for Human-oriented Communication (BARTHOC) of the applied computer science group at Bielefeld University.

Both robots are equally equipped with 41 actuators that enable human-like DoF for their movements. Thus, they can tilt and pan their heads to both sides and up and down, respectively. Besides, the complete head is movable in forward and backward direction. Furthermore, the eyes, each equipped with a color camera that have a resolution of 640×480 pixel (VGA), and a focal length of either 4, 6, or 8 mm, can be panned separately but tilted only for both eyes in the same angle. For the simulation of mimics, the artificial heads have actuators for the forehead, eye brows, cheeks, and the jaw as well. The torsos are equipped with additional actuators that allow shrugging the shoulders and a movement of the arms and hands while each hand is able to move its fingers separately. However, the hands are usable only for communicative gestures and not for manipulation tasks due to their construction.

With regard to the computational resources available for the BARTHOC robots, currently two stationary computers for each robot are used, one for the head and one for the torso. As the motor control unit and the cameras are separately connected with the computers, the resources can easily be extended if necessary.

After all relevant details for the hardware have been discussed, the software-based implementation of the proposed Object Attention System is described in detail in the following.

4.3 Data Representation for Intra- and Inter-module Communication

The related work for object learning showed that the communication between the human and a social robot needs to be supported by the Object Attention System in order to make the conversation as comfortable as possible. In order to be able to adapt to the rapidly changing challenges, a flexible data communication format has to be used as well. In the following, therefore, the communication scheme used is briefly pointed out. Additionally, the Object Attention System needs to be able to adapt itself to the varying environment. To face these requirements,

a global and fully verifiable setup configuration has been developed that is, subsequently, be presented. A detailed description on these issues is given in the appendix A.1 on page 119.

Flexible Data Exchange

The development of a cognitive operating Personal Robot demands for the incorporation of many components that communicate with each other. These issues lead to a very complex architecture and, thus, worldwide not a single research group is able to match every specific demand in the best possible manner. Hence, it is essential to provide interfaces for each module (e.g., for object recognition, speech processing, ...) that allow different research groups to interconnect their solutions as easily as possible. Because of this constraint, all interfaces no matter whether it concerns text-based or binary data are unified which offers for all modalities used a great deal of flexibility. The fact that the exchanged data embeds various types of formats, like, e.g., text in arbitrary languages or binary data (motor commands, sound, images, ...) additionally requires the capability to deal with these constraints using a neutral format container. In order to solve these challenges, only open, non-licensed, and internationally accepted standards are used for the different domains *Audio*, *Text*, and *Vision* within the Object Attention System. Furthermore, these standards need to be independent from a particular programming language or operating system which has been realized as well for the Object Attention System. This completes the brief discussion on the flexible data exchange issues as the details on the concrete approaches are described on page 119. The second aspect mentioned above concerning the need for a global configuration is, therefore, briefly given next.

Global Configuration of the Object Attention System

The proposed Object Attention System is designed to be integrated in various robotic systems. Thus, all environment-dependent settings, like names for communication channels or hardware parameters for cameras are adaptable without the need to edit the source code. In particular, a global and easy extendable XML-based configuration has been developed which is verifiable with an also developed corresponding XML Schema-based validation. This way, invalid entries are immediately displayed and the user is able to correct the values. While it is inconvenient under research conditions to always change the configuration file as soon as another robotic platform is used, the Object Attention System supports parameters which let the developer select the individual system-dependent configuration.

These selectable configurations always cover the same features that are divided into four semantically distinct blocks:

- Global configuration (locations of object and logging files)
- Communication connections
- Hardware setup (camera parameters, including position and orientation)
- Memory setup (size and storage duration of Short- and Long-Term Memory)

To sum up, a coarse introduction into the concepts used for the flexible interface of the Object Attention System has been given. These concepts support an easy integration of the Object Attention System in robots used by other research groups. Next, the temporal dependencies between the different modalities used are considered. This reflects the corresponding discussion in the related work of this chapter. Therefore, the developed Short-Term Memory approach is presented in the following.

4.4 Short-Term Memory

The developed Short-Term Memory optimizes the memory consumption, processing speed, and CPU load by its data-centralized and data-encapsulated character. Additionally, it is held flexible for the integration in other robotic systems. This becomes possible as the most important settings enable a quick adaptation of memory duration, memory size, and object storage location which is supported by the global configuration approach described above. As the Short-Term Memory contains all input data that needs to be accessed several times during one processing cycle of the Object Attention System, it offers an optimized representation for each modality.

In particular, the different interfaces are symbolically composited by the *Store* label in Figure 4.9. This interface is directly connected to the individual context-dependent memory structures for the processing of gestures, speech, person data, and object data. During each storage process, a label is attached that contains a *Best Before* timestamp for later validity checks. With this timestamp calculation and the subsequent storage in the appropriate memory structure, the overall storage process is complete.

As soon as all data is complete and the Object Attention System wants to access the memorized data, a structure called *Memory Synapse* is accessible by a separate interface. Within this structure, the actual temporal fusion of the modality data stored is done. During each access, the memorized data is checked for validity (age, completeness) first. If an error occurs, a meaningful error message is returned to the querying instance. Then, starting from the oldest stored values the temporal assignment of speech, gesture, person, and object data is performed in correspondence to the values that have been discussed on page 52 and the following ones. As an outcome, all relevant and temporally related data

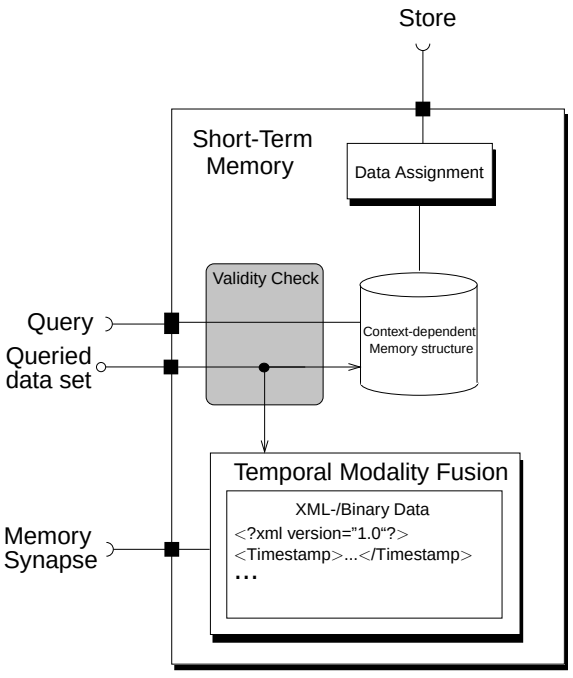


Figure 4.9: Schematic illustration of the Short-Term Memory and its interfaces.

is summarized in a structure called *OASObject* that integrates the attributes and values received by the gesture recognition, speech processing units, and the person tracking component. This OASObject is from now on the only global data structure needed for further processing tasks within one overall processing cycle (object learning or object recognition) of the Object Attention System. Thus, the efficiency mentioned above is achieved. After the fusion process is complete, is the OASObject accessible over a third interface category, in particular, the *Query* interfaces. They offer a generalized possibility to extract either the current OASObject or particular data set related to a specific modality.

For a more detailed illustration of the internal structure of the Short-Term Memory, please confer the description in appendix A.4 on page 126. As the basics of the Short-Term Memory have now been discussed, the important aspects of the object perception (visual, auditory) and representation is presented next. For clarification, this may not be confused with the OASObject as this is a data structure for internal use.

4.5 Perception and Representation of Objects

The perception and representation of objects referenced during an HRI presents besides the anchoring and fusion process of sensor signals to appropriate symbolic symbols the most challenging task for the development of a multi-modal Object Attention System. Therefore, a special focus has been layed on the visual learning and representation algorithms for a priori unknown object instances. Therefore, the outline of this section is as follows.

At first the necessary conversion between symbolic and numeric expressions is presented. Then, the computation for the actual location where the robot refers to, the Region-Of-Interest, is described. After that, the visual segmentation and representation used for the final object views are pointed out. Finally, the corresponding ontological object representation in a spatial and a textual semantic sense is regarded.

4.5.1 Modality Converter

The visual exploration of a scene utilizes common image processing algorithms. Those algorithms use numeric values, e.g., for the representation of colors denoted in a specific color space, like the *Red Green Blue (RGB)* or *Hue Saturation Value (HSV)* color model. However, the user of a robot does not have an interest in the learning of thousands of numerical color value combinations, instead he wants to deal with colors in a familiar way. In detail this means that he wants to use symbolic color names, like 'red' or 'cyan'. In order to provide such a symbolic $\Leftrightarrow$ numeric transformation, the Modality Converter has been developed for the Object Attention System, cf. page 121 for details.

The Modality Converter is designed as a stand-alone application as well. For reasons of flexibility, its access interfaces are fully based on XML which enables other modules, like the distributed Long-Term Memory, to state queries to the Modality Converter. In its current implementation, the Modality Converter supports the following four features for transformation tasks:

- Color
- Relation
- Size
- Shape

Except for the two feature types 'Relation' and 'Shape', all other transformations are performed in a classical mapping task symbolic $\Leftrightarrow$ numeric. Nevertheless, they all use a database-like model of a lookup table which is encoded in XML as well. This enables several advantages, like easy adaptability and automatic verifiable entries.

The basic conversion scheme is explained by the following simplified lookup table 4.1. In the most left column, the *Predicate* name, e.g., *color* is given, which corresponds to the symbolic expression used by the dialog, speech understanding and speech recognition component. All these predicates have values, e.g., *red*, as shown in the second column. This is enough information to enable the Modality Converter a conversion into a format usable for the Object Attention System. Depending on the predicate used, this destination format varies a lot, as the *Numeric/Model* column exemplifies.

Predicate	**Symbolic value**	**Numeric / Model**				
Color	red	Model	Ch.	Ch. 1	Ch. 2	Ch. 3
		HSV	6	0...4	25...29	151...155
Relation	Obj_1 under Obj_2	$Obj_1.y < Obj_2.y$				
Size	small	5...12				
Shape	round	Haralick				

Table 4.1: Simplified lookup table of the Modality Converter.

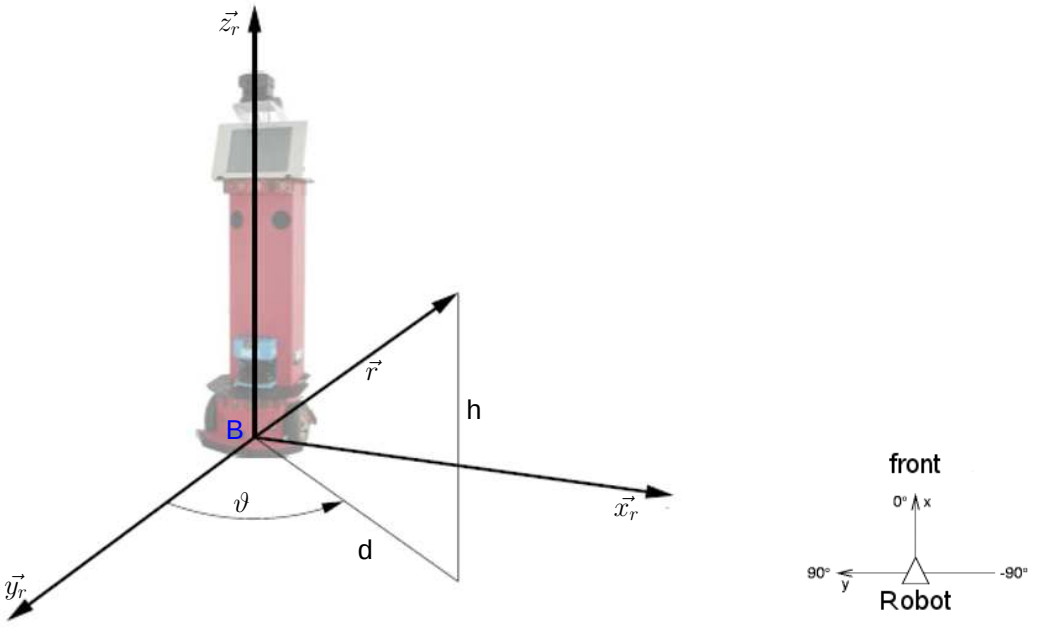
From this structure arise three possible conversion situations. First, all needed entries for symbolic and numeric values are specified. In this case, the Modality Converter returns the appropriate answer to the querying module. Secondly, in some cases, .e.g., a symbolic name, like "transparent" does not have a corresponding numeric value defined in the color space used. Thirdly, no numeric value exists for a given symbolic name yet. In the latter case, although not yet implemented, the Object Attention System could resend the color values of the determined Region-Of-Interest, e.g., at the position of its center of mass. That way, the lookup table of the Modality Converter can be completed and, thus, a previously unknown symbolic color can subsequently be used by the Object Attention System.

The two strategies used for the features *Relation* and *Shape* follow the same principles. Both map the large variety of possible predicate names to a relatively small set of mathematical expressions (e.g., $<, >, =$) or names for shape recognition approaches [WC05, Ber05] (e.g., Haralick, Least Median of Squares, Fourier Descriptors, or Minimum Bounding Rectangle), respectively. Thus, the user can iteratively optimize the extraction of an object view without the need to know about the particular underlying mathematical methods.

After the modality conversion is completed, the Object Attention System is able to use the transformed values to focus the camera on the referenced Region-Of-Interest.

4.5.2 Determination of the Region-Of-Interest

The alignment of the camera on the referenced Region-Of-Interest which ideally contains the object is a very challenging task. In order to solve this task, two different algorithms were implemented in the Object Attention System, both mainly using the output of the gesture recognition system. The first one for the principally 2D-gesture recognition in combination with the depth data provided by the laser range finder of the robot BIRON. The second algorithm is able to deal with the data provided by the 3D-Body Model Tracker in combination with the gesture recognition, cf. section 3.1 on page 36. Thus, the latter one is a great deal more precise than the first algorithm based on the gesture recognition using skin-color region tracking. Nevertheless, the goal for both approaches consists in the alignment of the object camera which uses the robot coordinate system depicted in Figure 4.10. This is of special interest, as the robot and the object camera, respectively, use a completely different coordinate system than the gesture recognition, although all positions, finally, need to be transformed into the object camera coordinate system.



(a) Robot's cylindric coordinate system, adapted from [Lan05].

(b) View from top on robot's coordinate system.

Figure 4.10: Robot coordinate system used for the Object Attention System.

As Figure 4.10(a) shows is the object camera coordinate system a cylindric one with its origin B for the height $\vec{z_r}$ on the bottom of the robot. In that plane, the axis $\vec{y_r}$ pointing to the front origins in the middle of the robot, while the orthogonal axis $\vec{x_r}$ radiant from the axis of the drive wheels. Furthermore, the horizontal angle ϑ

[rad] has its 0° angle pointed to the front of the robot, measured counterclockwise as illustrated in Figure 4.10(b). The remaining components height h [m], distance d [m], and radius r [m] have the same origin point B as the Cartesian $\vec{x_r}, \vec{y_r}, \vec{z_r}$ coordinate system as shown in Figure 4.10(a).

The computation of the position for the Region-Of-Interest starts in both cases as soon as the dialog component sends the order to align the camera on an object. For efficiency reasons, the following computation is only performed for the relevant gesture and not for all stored gestures in the Short-Term Memory. This becomes available by the analysis of the temporal correlation between speech and gesture, as it is described in section 4.1. Based on the content of the XML data sent by the gesture recognition module, the Object Attention System decides how to proceed. In particular, two cases are distinguished, one for 2D-data, and one for 3D-data.

Dealing with 2D-Gesture Data for the Region-Of-Interest

If the 3D-Body Model Tracking System is not used, the gesture recognition itself supports the Object Attention System with the 2D-hand position and an estimated 2D-pointing direction α (0°...360°). Thus, it is necessary for the Object Attention System to extrapolate the pointing direction as otherwise the center of mass of the hand would be interpreted as the center of the Region-Of-Interest. The following paragraph, therefore, describes the basic proceeding in a simplified manner of the algorithm applied to overcome this restriction.

To get an appropriate coordinate for the Region-Of-Interest, trigonometric functions for right triangles are used to cover the various pointing directions, see Figure 4.11.

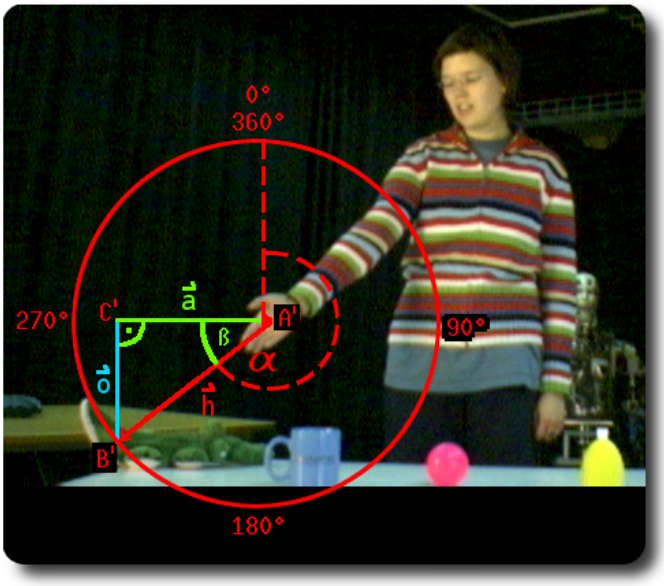


Figure 4.11: Illustration of geometric 2D-approximation for the Region-Of-Interest.

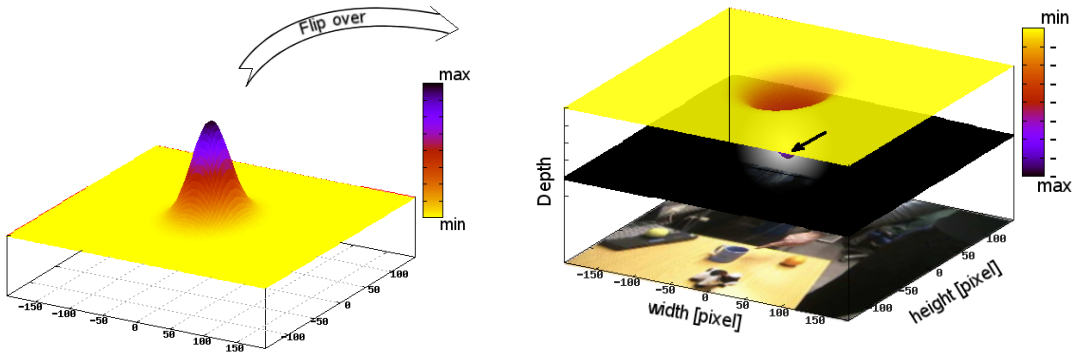
In particular, the slope m of the red-colored hypotenuse $\vec{h}$ is calculated which runs parallel to the pointing direction, cf. equation 4.1.

$$m = \frac{\vec{o}}{\vec{a}} = tan\beta \tag{4.1}$$

The hypotenuse is then, starting from Point A', continued in pointing direction for approximately 30 cm. In this way, the center for the Region-Of-Interest (B') is determined. The shift of the Region-Of-Interest for about 30 cm is indeed a very coarse approximation, but in real experiments it has been proven as accurate enough.

The Object Attention System so far does not have an estimated distance (Robot $\Leftrightarrow$ Hand) for the gesture position. Thus, the distance value of the person provided by the Person Tracking and Attention System [Kle05, Lan05, Spe05] is used. However, it has been shown that if the user's hand is in the upper half of the image captured by the gesture camera, the referenced location is a great deal nearer to the camera than the legs. This is caused by the circumstance that in such a case the user usually points to a location on a tabletop in front of him. In order to compensate this discrepancy, the distance value is then simply divided by the factor 2.

As a result of the calculation of the Region-Of-Interest, a gesture-based Attention Map, like it is depicted in Figure 4.12 is generated.



(a) Gaussian distribution.

(b) Gesture-based *Attention Map* (isometric view) modeled with a Gaussian distribution.

Figure 4.12: Visualization of a gesture-based Attention Map.

In particular, the Figure shows the input image provided by the object camera, which is overlayed by a black hidden layer and the actual gesture-based Attention Map modeled as Gaussian distribution. It illustrates the effect of the gesture map. In the center of the distribution peak (marked by the black arrow), the black hidden layer is completely penetrated while with decreasing amplitude of the *Gaussian* distribution $G(x, y)$, the penetration decreases as well. The mathematical relation for $G(x, y)$ is described in equation 4.2. The variance σ^2 is regulated with the mean size value provided by the Modality Converter. The black area of the Attention Map causes a fade out of the image parts not belonging to the Region-Of-Interest.

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \tag{4.2}$$

This method to calculate the location for the Region-Of-Interest is indeed only a coarse estimation. Therefore, the 3D-based estimation of the Region-Of-Interest

has been implemented as well which in turn applies the integrated 3D-Body Model Tracking System. The following section, hence, describes the algorithm developed for the Object Attention System that allows a more precise localization for the Region-Of-Interest.

Dealing with 3D-Gesture Data for the Region-Of-Interest

The preliminary evaluation of the 3D-Body Model Tracking-based gesture information has shown that it provides a great deal more accurate positions of the estimated Regions-Of-Interest than the 2D-based gesture recognition does. Therefore, the localization algorithm for the Region-Of-Interest has been adapted in order to be able to automatically deal with the changed preconditions. Nevertheless, it has to be noted that the usage of the 3D-Body Model Tracking System is currently not capable to process the image streams in an online Human-Robot Interaction scenario, as the 3D-Body Tracker is still not optimized.

However, before the actual algorithm is explained, an overview of the different coordinate systems used for the localization task is given first. The destination coordinate system still remains the one for the object camera, as depicted in Figure 4.10 on page 64. Furthermore, the gesture recognition and the body tracker, respectively, use an orthogonal 3D-Cartesian coordinate system as the right half of Figure 4.13 illustrates. Besides these two already known coordinate systems, an additional spherical user coordinate system is used. Although it would not be necessary to introduce a third coordinate system from a mathematical point of view, it is helpful for an easier illustration and implementation of the algorithm. This user coordinate system has its origin C at the *HEADPOS* location provided by the 3D-Body Tracking System, described in section 3.1 on page 36. While the main directions are represented by the vectors $\vec{\xi}$, $\vec{\psi}$, and $\vec{\zeta}$, are the horizontal and the vertical angle denoted by φ and ϑ, respectively. This additional coordinate system is primarily used to enable an easy extrapolation of the 3D-pointing direction which is finally used to localize the Region-Of-Interest. In the following paragraph the algorithm is explained in detail.

In a first processing step, the positions for the *RIGHTFINGERTIP* D (marked as red square in Figure 4.13) are extracted. If they are not available, the also red marked *RIGHTHANDPOS* E tags are evaluated instead. As the latter one marks the wrist, a correction distance of 15 cm is added in motion direction of the pointing gesture. Usually three different locations of the *RIGHTHANDPOS*, *RIGHTFINGERTIP*, and the *HEADPOS* are present, due to the last three timesteps of the *Condensation-based Trajectory Recognition (CTR)* (cf. page 38). Consequently, the last 200 ms ($3 \cdot 66{,}6$ ms, related to the 15 fps used for the CTR) are considered and if they are sent to the Object Attention System, their mean value is calculated in order to suppress too large variations for the positions. This becomes necessary, as the 3D-Body Tracker uses non-deterministic probabilistic methods which result in recognition errors, especially in the depth values, as described by Schmidt in [SKF06]. These errors are mainly caused by the monocular camera used, because with only one point of view, a lot ambiguities occur for different poses. A motion model that could reduce these ambiguities is, currently, not yet implemented in the 3D-Body Tracking System.

The second processing step transforms the extracted values related to the body tracker coordinate system which is spanned by the axes $\vec{x}$, $\vec{y}$, and $\vec{z}$ into the cylin-

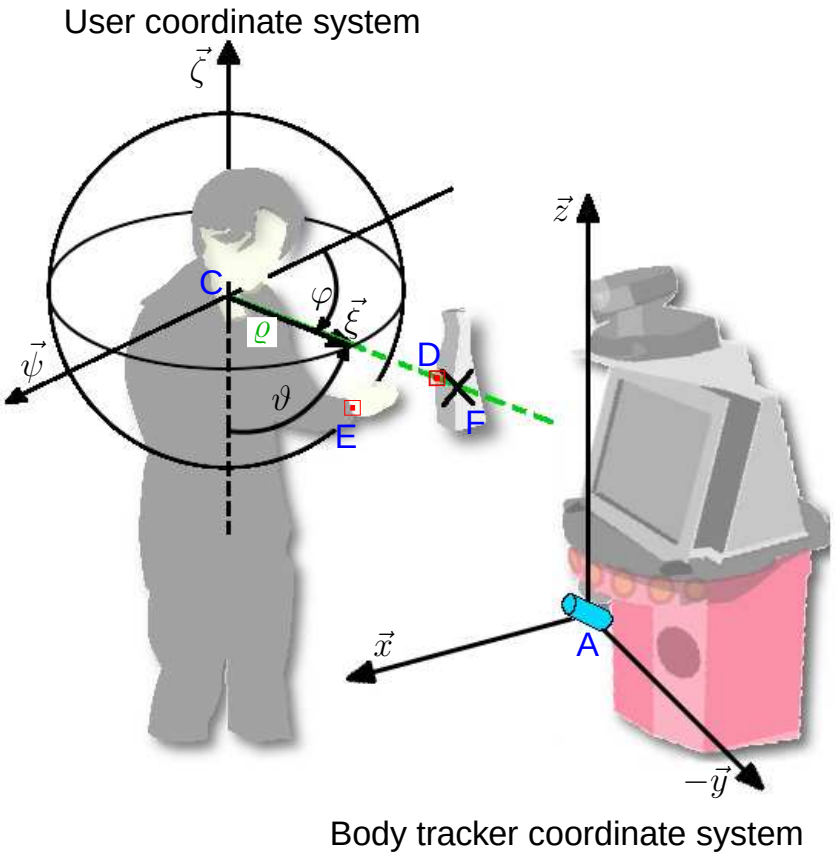


Figure 4.13: 3D-Coordinate Systems used for the Object Attention System.

dric object camera/robot coordinate system (Figure 4.10). This transformation considers the horizontal and vertical offset as well as the tilt of the gesture camera related to the object camera coordinate system.

Next, the different user-dependent positions are transformed into the spherical user coordinate system. This enables a simple distance adaption between the user's hand position and the center of the referenced Region-Of-Interest. This distance, however, is not fixed and depends on the object size verbally specified by the user. The corresponding numeric size value is returned by the Modality Converter as described in section 4.5.1 on page 62. The same size value is used for two other aspects. First, it is used to generate an *Attention Map* in a similar calculation as it is described in the section before for the 2D-gesture processing. Secondly, the size value is used to adjust the zoom factor of the object camera in order to capture a more detailed view of the Region-Of-Interest. Finally, after the position and the size of the referenced location has been determined, a last transformation outgoing from the user coordinate system into the object camera coordinate system is calculated. Consequently, the object camera is then aligned to the computed location of the Region-Of-Interest while the calculated zoom position is also transmitted to the camera. During the alignment of position and zoom, the actual camera settings are continuously evaluated with a frequency of approximately 10 Hz. Thus, the Object Attention System is able to initiate further processing as it is ensured that the camera has reached its final position. In the case of an error, the Object Attention System sends a message to the dialog component that an alignment on the object is not possible.

For reasons of an easy evaluation, another unit test has been developed. This is discussed later on in the evaluation chapter. Assuming that the camera has successfully been aligned towards the Region-Of-Interest, a visual scene analysis can be performed. The methods of analysis used by the Object Attention System, therefore, are the topic of the next paragraph.

4.5.3 Visual Object Representation

The visual object representation and the generation of object models as abstract as possible that allow a robust recognition in an unknown and cluttered scene is one of the most difficult challenges for the development of the proposed Object Attention System. That this is still an unsolved problem has been shown in the related work to this chapter on page 47 as well as in chapter 2. However, the overview has pointed out that an intelligent reduction of all input image data is essential for the establishment of a successful Object Attention. Thus, some of the most promising approaches for data reduction have been considered in the proposed Object Attention System of this thesis. Regarding the visual representation of objects, this reduction is, therefore, based on the features *Color*, *Depth*, *Relations*, and *Gestures*. The feature *Shape* is also very helpful, but due to clutter, occlusions, varying lighting conditions and other environmental influences difficult to deal with. Nevertheless, a couple of intelligent approaches already exist for shape-based object recognition, e.g., [CD02, LLS04, BBM05]. In this work, however, this feature could only partially be considered due to the complex algorithms needed, as described in section 4.5.1 on page 123. At first a data reduction is done by the gesture-based restricted selection of an appropriate Region-Of-Interest, like it is described in the section before. Then, a detailed image analysis follows which is described next.

(a) Original input image. (b) Smoothed input image. (c) Attention Map for “Blue".

Figure 4.14: Demonstration of color attention. Images taken from [Sae05].

Before the actual visual object appearance within the Region-Of-Interest is extracted, some image preprocessing is done in order to improve the results. This preprocessing is performed on the original input image 4.14(a), like it is captured by the object camera. As the camera used has a *CCD (Charged Coupled Device)* sensor, the image contains noise due to physical effects, like *thermal noise* or *dark current*. Unfortunately, these issues have a greater impact in indoor scenarios as a longer exposure becomes necessary. Thus, in order to reduce the influence of the noise, a Gaussian filter mask is convolved with the underlying original image. The result of the convolution is shown in Figure 4.14(b) while in this example a mask with a size of 11×11 pixel has been applied. This mask size has been

proven as a good compromise during the evaluation phase of the Object Attention System. The filter is optional, integrated as an *iceWing* [Löm06] plugin while it uses algorithms of the OpenCV library (cf. section 4.3 on page 121). Besides the configurable *Gaussian filter*, additionally *Bilateral*, *Blur*, and *Median* filters can be applied. In this way, the preprocessing can be optimized for the relative task.

In the following the actual view-dependent object extraction algorithms are described. Therefore, the color-based generation of Attention Maps is described next.

Color-based Object Analysis

The subsequent processing for a *Color Attention* is realized as a configurable plugin as well, which has been developed in cooperation with M. Saerbeck in his master thesis [Sae05]. Although the color analysis is in principle also optional for the Object Attention System, it is useful to always consider the color, as otherwise the resulting learned object view will probably contain more background image parts. In the end this will usually produce a less accurate result. To overcome this constraint, a depth-based Attention Map has been developed together with M. Köllmann [Köl06] for the Object Attention System that, however, cannot be used simultaneously with the color-based approach so far.

The visual routines in turn cause the Object Attention System first to send a query to the Modality Converter in order to get the symbolic color name transformed into numeric values. After a successful query, a color-based Attention Map is created which highlights only the specified color in the image, like it is shown in Figure 4.14. The calculation of the Attention Map in turn contains several processing steps. In the following, only the *HSV* color space is regarded as its use resulted in the most promising Attention Map while the colorspaces *RGB*, *LUV*, *YUV*, and additionally *GRAY* images have been considered as well with a corresponding conversion plugin. In particular, the most important advantage of the HSV color space is the separation of color and intensity processing that enables more robustness against varying lighting conditions as, e.g., the RGB color space. This can easily be verified in the image 4.14(c), since all blue objects are highlighted, although their colors have different intensities (e.g., trash bin vs. cup). However, the plugin used for Color Attention is independent from the color space used. Only the calculation of the attention value for each image pixel is identical, no matter what color space is used. This calculation is described next.

The equation 4.3 used for an Attention Map A on an image I is described below. The overall Attention Map is defined by the sum of activations $A_{i,j}$ that are calculated for each pixel of an image I with its dimensions $dim(I) = X, Y$. This relation is subsequently represented as function $\gamma(z)$, where z is a geometric mean value used for noise reduction.

$$A = \sum_{i,j}^{X,Y} A_{i,j} = \gamma(z) \tag{4.3}$$

Now in order to suppress outliers, the geometric mean value z is defined on a neighborhood N for each position i, j. The resulting value for each neighborhood

position is then compared to the reference color value c which enables the measurement of the similarity between the reference value and its surrounding pixel. The best results for an appropriate activation have been shown with the empirically determined equation 4.4. The variable d denotes the maximum distance of two colors related to one color channel. For an 8 bit color image, like it is used for the Object Attention System this is consequently the value 255. The exponent p determines the slope parameter of the function.

$$\gamma(z) = 1 - \left(\frac{z}{d}\right)^p = 1 - \left(\frac{\sqrt{\sum_{m,n}^{N}(I(i+m,j+n)-c)^2}}{d}\right)^p \tag{4.4}$$

For visualization, the Figure 4.15 presents the equation applied with different parametric values of $\frac{1}{2}$, 1, and 2 for the slope parameter p and a value of 255 for the color distance d.

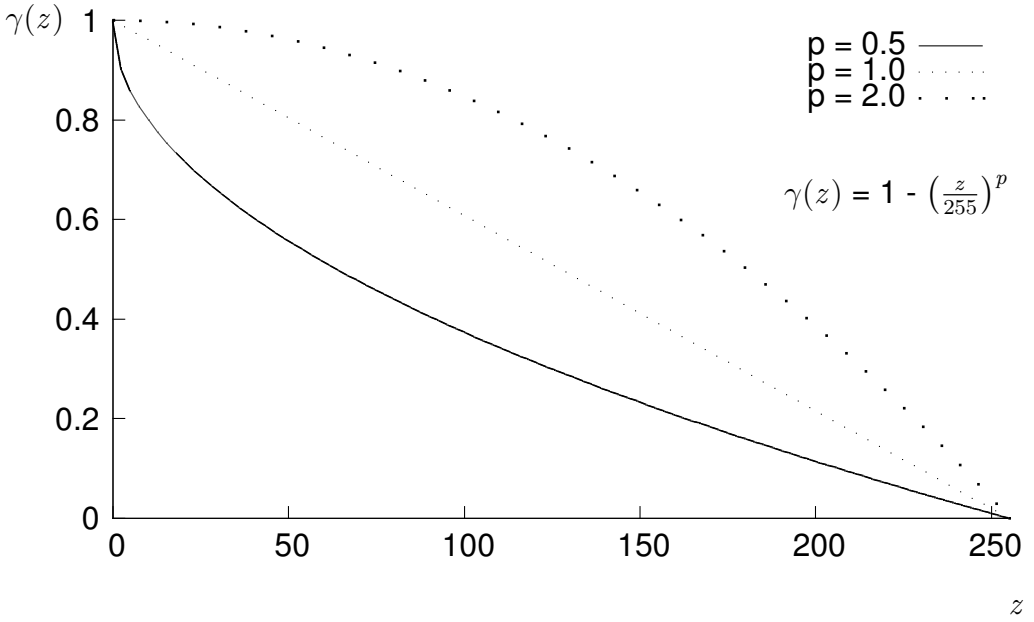


Figure 4.15: Activation function used for a color-based *Attention Map*.

The different curves illustrate that pixel which are similar to the reference color value get a small mean value z and, thus, a high activation. On the contrary, the more different the color value of a pixel is, in comparison to the reference color, the lower is its activation. In order to be able to evaluate verbal expressions, like “reddish” or “ginger”, the slope parameter p has been proven as very helpful. Its effect can be described as follows. If it is less than 1, the function value decreases faster than for a slope parameter which is greater than 1. This reflects the intuitively assumed relation as a slope regulator. A special case represents a slope value of 1, which causes a linear dependency between the color activation and the color similarity. As an outcome of the color-related Attention Map, an object view can more easily be extracted.

Although the color-based Attention Map is mostly sufficient, a depth-based Attention Map has been integrated as well for the Object Attention System in order

to get an improved object view of the object regarded. This depth evaluation is described next.

Depth-based Object Analysis

The development of robots can be divided into two main directions. Those that are designed with a focus on a technical appearance which provide all kind of sensors that are not comparable with the human-like perception, like ultra-sonic sensors or laser range finders. The second category represents anthropomorphic robots that are equipped with human-like sensors, like two cameras for the eyes, or two microphones for the ears and so on. In order to cover the latter category as well, the above mentioned model for depth-perception based on a stereo view is described together with its underlying algorithms in the following.

To uphold the flexible character of the Object Attention System, the depth-related attention processing has been implemented as a separate iceWing plugin, too. Thus, it is optional, like the preprocessing or color-related plugins described above.

The algorithms applied for the depth-based acquisition of object positions support a specific human-like feature, in particular, they are designed to use a flexible stereo camera head where its two cameras can be panned and tilted even in relation to each other. Furthermore, a second requirement is met that offers the possibility for a relatively fast depth estimation during a Human-Robot Interaction. Therefore, the principle calculation of a depth-based Attention Map is divided into three processing steps:

- Intrinsic calibration for each camera
- Extrinsic calibration for the relation between the two cameras
- Determination of a disparity-based *Depth Map*

The basic algorithms used for each of these processing steps is briefly presented in the next paragraphs. For a more detailed description, the referenced literature needs to be considered.

Intrinsic Camera Calibration

The intrinsic camera calibration becomes necessary for a perspectively correct reconstruction of locations in the scene. In this case, the calibration is based on the algorithm proposed by Hartley in [Har94a, Har94b] that analyzes the projective distortion in images. The intrinsic calibration begins with two images captured by one camera with a slightly different camera orientation. Thus, the algorithm can compare these images for the distortions. For these two images, a *Homography* is calculated that describes the transformation between the subsequently captured images best by minimization of the squared error between these two images. However, due to wrong assignments for pairs of image points, additionally the *RANSAC* (Random Sample Consensus) algorithm, described by Fischler and Bolles [FB81] is applied in order to reduce the squared error. After the Homography is known, the actual camera parameters (position and orientation) can be calculated. Finally, a median filter is applied on the parameter values which

reduces errors during the determination of the calibration matrix K which is described in equation 4.5. Here, the variables k_x and k_y describe the magnification in x and y direction. The variables p_x and p_y describe the principal point of the image and s is a skew parameter which corresponds to a skewing of the coordinate axes used.

$$K = \begin{bmatrix} k_x & s & p_x \\ 0 & k_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.5}$$

After the camera calibration matrix for each camera is calculated, the extrinsic camera calibration takes place.

Extrinsic Camera Calibration

The relation of orientation for the two cameras of an anthropomorphic robot is calculated by the extrinsic camera calibration. In the model used, this calibration is separated into two parts. First, the calculation of the camera angles, and secondly the determination of position and orientation related to each of the cameras and the head of the robot, respectively. The second part uses the epipolar geometry, like it is described by Hartley and Zisserman in [HZ04]. At the end, the performed calculations result in two camera matrices C_l and C_r for the left and the right camera, see equation 4.6. The variable K_i represents the two calibration matrices from the intrinsic calibration, while R_i describes the rotation between the cameras and $\vec{t}_i$ denotes the translation between the left and the right camera related to a common coordinate system.

$$C_i = K_i[R_i|\vec{t}_i \quad \text{with } i \in \{l, r\} \tag{4.6}$$

As a final processing step, a depth value can be calculated for each position in the image. However, this needs to be qualified, as this method works properly only on locations with textures.

Determination of Depth Values

The depth value for a given scene object becomes available after the cameras have been calibrated and a depth image can be calculated. The model used for the Object Attention System in order to create a depth image uses triangulation-based algorithms. In particular, a disparity image is calculated according to the algorithm proposed by Birchfield and Tomasi in [BT98]. However, the disparity can only be calculated if corresponding image parts can be determined in both images, the one of the left camera and the one of the right camera. Here, the necessary feature points are extracted with help of the PCA-based SIFT features [KS04] introduced by Ke and Sukthankar. They have proven as very robust against translations, rotations, and different scaling of regarded scenes. An example for a disparity image is shown in Figure 4.16(a). This disparity image has been calculated from the images of the left (Figure 4.16(b)) and the right camera (Figure 4.16(c)). The different gray values in the disparity image present the distance of the objects. The brighter it is, the more different is that location in the two cameras and, hence, the nearer is the object located to the camera. The

(a) Disparity image: Bright ≡ near; Dark ≡ far or disparity not calculable.

(b) Image of the left camera.

(c) Image of the right camera.

Figure 4.16: Illustration of a disparity image with two objects. The images have been taken from [Köl06].

horizontal stripes are caused by the linewise computation of the disparity image which can partly be corrected with appropriate smoothing techniques.

Summarizing, two powerful segmentation approaches have been developed for the Object Attention System. Nevertheless, such a segmented region often represents an object only partial. Hence, the logical conclusion is to combine various image parts to a semantically combined set of image patches in order to approximate a complete object view. As described in the related work to this chapter offer feature graphs a fine solution for this problem.

Graph-based Object Representation

The graph-based approach which has been developed for the Object Attention System in the diploma thesis of M. Saerbeck [Sae05] combines different coordinate systems for an appropriate object representation. This can be described best on the example depicted in Figure 4.17.

For the cube shown in Figure 4.17(a), two different colored sides (red, cyan) should be combined. Therefore, the centers of mass (CoM) have been determined for the red-colored side R_1, the cyan-colored side R_2, and for the object itself S. The resulting coordinates are located in the original image coordinate system. In a second processing step, these coordinates are transformed into an object-centered coordinate system with the points R_1' and R_2' at the locations of the center of mass for each feature (Figure 4.17(b)). This way, two new feature coordinate systems can be created with the resulting locations R_1' and R_2' as principal points, shown in Figure 4.17(c). Finally, a new image coordinate system is defined with its principal point matching the origin of the relative feature coordinate systems (Figure 4.17(d)). This enables the Object Attention System to generate a feature knot that contains the following information:

- Type of the feature, e.g., color
- Value of the feature, e.g., red
- Position, relative to the object center of mass (x, y)

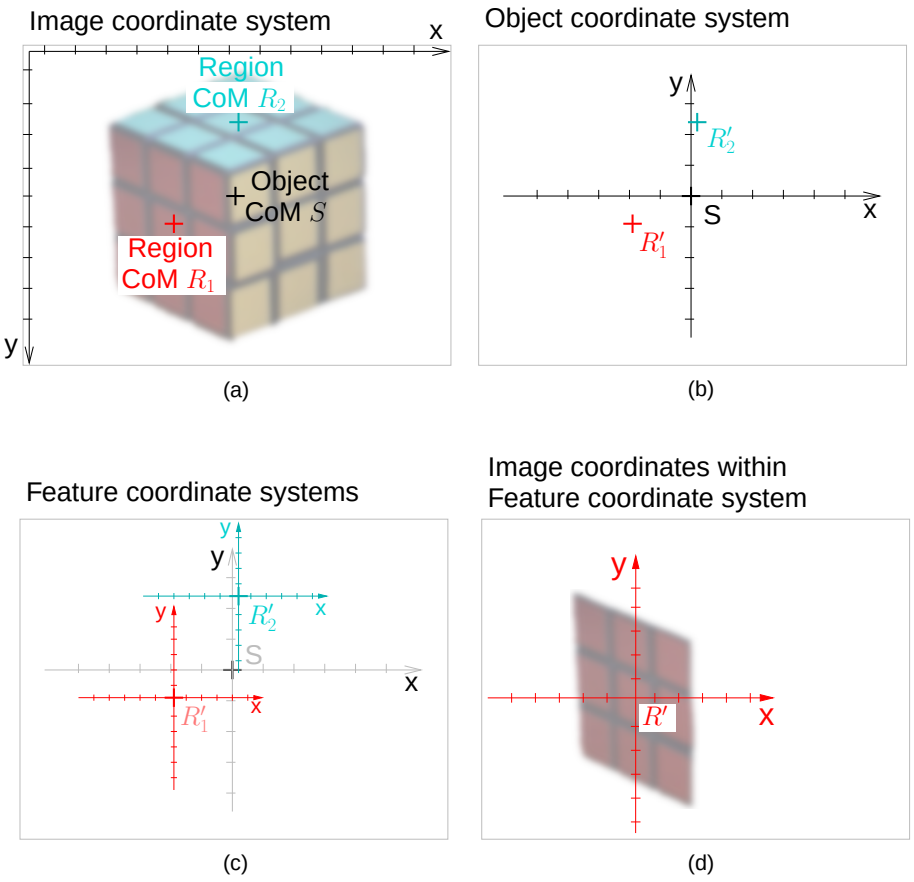


Figure 4.17: Coordinate transformations for graph-based object representation. The images have been adapted from [Sae05].

- Amount of pixel, selected by one feature

In this case, each pixel is a tuple of position (x, y), color value g, and activation value a : $M = \{(x, y, g, a), \ldots\}$. Every newly generated feature knot is then inserted in a feature graph as depicted in Figure 4.18.

As the image illustrates, the different feature knots define a feature-graph which describes the relations between each single feature to another feature. The edges of the feature-graph are designed as distance vectors. Due to this relative position information, the spatial relations between all feature knots are well-defined for an object. Furthermore, this enables an incremental learning of an object, while it is possible either to add new feature knots or to remove obsolete feature knots from one object representation.

The presented graph-based solution for extracting object views is indeed very powerful, but makes it necessary for the user to refine an object view several times if it consists of more than one color. Hence, alternatives have been tried out as well that might reduce the amount of necessary interactions to learn an appropriate single object view. An approach that has been applied is briefly presented in the following.

Alternative Image Segmentation

The proposed object segmentation is sometimes not sufficient, especially if the pointing gesture is of poor quality or the object color has not been specified by

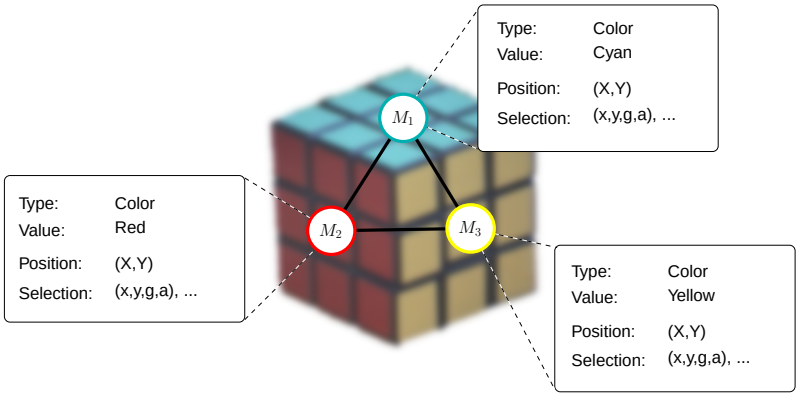
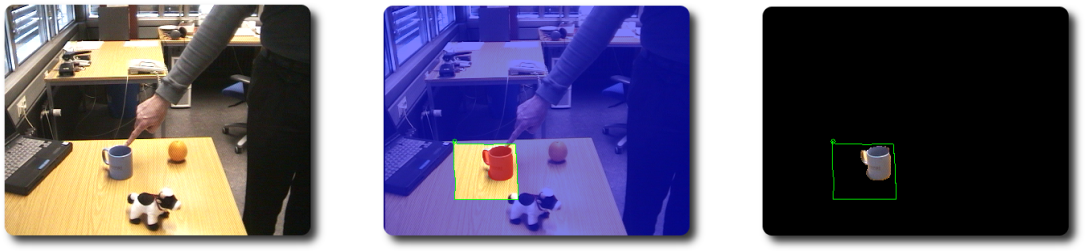


Figure 4.18: Object-centered feature-graph with three exemplary color features. The image has been adapted from [Sae05].

the user. For this case, an experimental implementation of the segmentation algorithm *GrabCut* which has been introduced by Blake et al. in [BRB+04, RKB04] has been considered as well. As it goes beyond the scope of this thesis, the algorithm is not explained because it has only been used. Nevertheless, the experiments showed for some environments promising results which is depicted in Figure 4.19. In detail, the Figure 4.19(a) shows the initial scene image without a selected Region-Of-Interest. The finding of a region is, therefore, illustrated in Figure 4.19(b), while the Region-Of-Interest has been manually selected. As a result, the Figure 4.19(c) demonstrates a fine extracted object view.



(a) Original scene with a user pointing to an object.

(b) Manually selected Region-Of-Interest.

(c) Segmented object view.

Figure 4.19: Illustration of the applied GrabCut algorithm.

However, the experimental result that is presented here took a lot of trials and even more time to calculate. The underlying algorithm is mainly based on the iterative energy minimization and uses a couple of probabilistic approaches. This results in a non-deterministic behavior that is, furthermore, occasionally very time-consuming ($>$ 1 minute on a desktop computer @2,4 GHz with 512 MB RAM). Hence, the approach has been considered as practically not usable for a natural Human-Robot Interaction which demands for relative short response behaviors of the robot. Additionally, even the GrabCut algorithm does often segment an object partially. For these reasons it has been determined, that the color-, depth-, and graph-based approaches are more suitable for a convenient interaction. Nevertheless, as especially for fine-grained selections of an appropriate Region-Of-Interest, the gesture recognition used is often too coarse, a Graphical User Interface has been integrated to overcome these constraints.

4.5.4 Graphical User Interface

The FLTK-based [FLT06] Graphical User Interface [Sae05] for the Object Attention System, developed in cooperation with M. Saerbeck provides a couple of useful interaction enhancements as it is described on page 42 and the following ones.

In this section, the focus lies on the visual representation of an object, like it is described in the paragraphs before. The image 4.20 shows the view for updating an already learned object. The GUI informs the user how many objects have successfully been learned during the ongoing interaction and allows him with the slider to select the object that he wants to update. The right half of the window illustrates which symbolic information about the object can be updated.

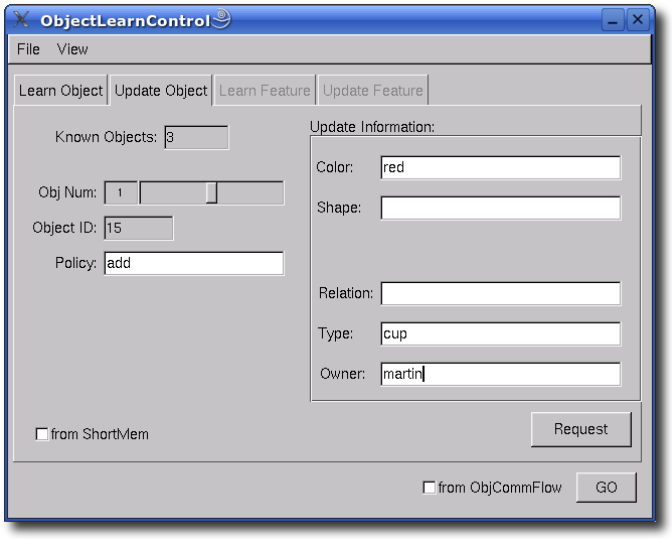


Figure 4.20: Graphical user interface used for written commands. Cf. text for a detailed description. The image has been taken from [Sae05].

To demonstrate an exemplary learning of a multi-colored object, a typical scenario is depicted in Figure 4.21. The figure shows the original input image containing several objects in Figure 4.21(a) from that the greenish marked blue/red tape should be learned. During a first scene analysis, all red-colored areas are extracted within the Region-Of-Interest (Figure 4.21(b)). The black parts in the image represent the non-red colored areas which are internally tagged as transparent. Subsequently, all blue-colored areas within the selected region are extracted as well as shown in Figure 4.21(c). By fusing these two colored views into a single object representation, an object view suitable for later object recognition tasks is created, which is depicted in Figure 4.21(d).

Summarizing the visual object representation, a couple of approaches have been presented that support features which can easily been verbalized, like *Color* and *Relations*. Nonetheless, the chosen representation can be used for other features as well, like *Object Type* or *Owner*.

Besides the visual object representation, a representation formalism has been implemented that is able to assign a sound to an object which is described next.

(a) Evaluation scene with different colored objects.

(b) Red parts of a multi-colored object.

(c) Blue parts of a multi-colored object.

(d) Fused red- and blue-colored object regions.

Figure 4.21: Representation of a multi-colored object.
The images (b,c,d) contain the learned object parts within the greenish marked Region-Of-Interest in image (a). The images have been taken from [Sae05].

4.5.5 Sound Collector

The audio signal provides valuable information in an Human-Robot Interaction. It serves as input for the speech processing units and can contain the sound of an object as well, e.g., the tone of an alarm clock or the barking of a dog. In order to capture this sound, the *Sound Collector* has been developed for the Object Attention System. Like all other input cues, it is realized as a stand-alone application which is fully accessible via XML documents. Thus, not only the paradigm of the distributed architecture for the Object Attention System is supported, but it also allows other modules, for instance, the dialog module to access and use the Sound Collector. This enables the robot to react on certain sounds and additionally supports applications, like a personal message service in similarity to a computer log or answering service for people.

The Sound Collector uses capabilities of the *Environment for Statistical Model Estimation and Recognition on Arbitrary Linear Data Arrays (ESMERALDA)* developed by G. Fink [Fin99] which is applied for the speech recognition as well. While sound is already recorded for speech analysis, it is possible to permanently store captured sounds as separate audio files. In particular, every time when the user talks to the robot, the Person Tracking and Attention module (see section 3.1 on page 34) activates the sound recording in order to enable the speech recognition system and, hence, to analyze the user's utterance. This is illustrated in Figure 4.22. The Sound Collector benefits from this situation as it is able to search for a particular audio file by interpreting its last modification timestamp t.

To extract a specific object sound, all recorded audio files that are newer than the sound file which contains the utterance with a specific keyword, like "sound", are considered, cf. the green marked time span in Figure 4.22. Then, the oldest file ending at time x and that has been modified after the queried timestamp t is selected, but only if it is within a specified time span of 30 minutes at maximum. This time span supports two functionalities and has been empirically determined. On the one hand it ensures that the selected sound or utterance, respectively, probably contains the expected object sound, because it is unlikely that a sound which ends more than 30 minutes after the queried timestamp t still belongs to an object sound or a speech memo. On the other hand this threshold limits the maximum recording time for a single audio file and, consequently, the file size

which enables a more reactive system whenever the data is transported within the active memory infrastructure.

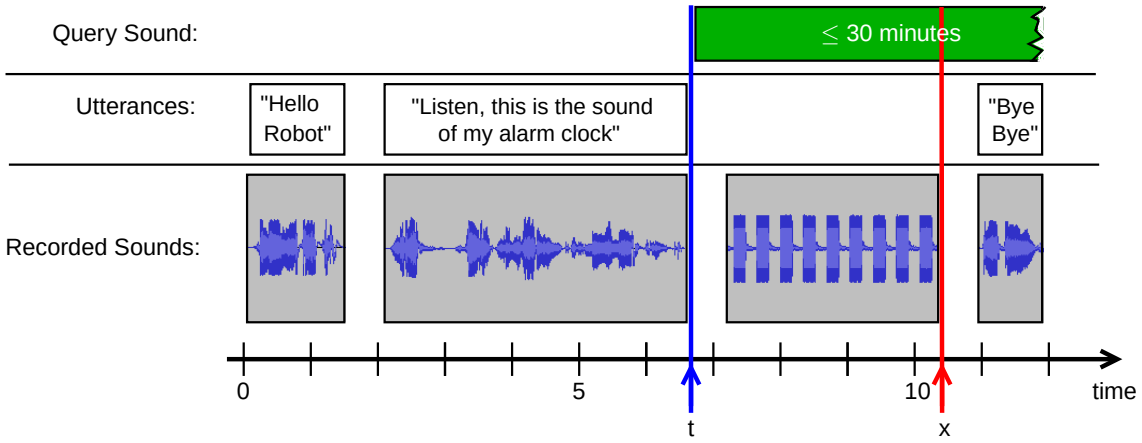


Figure 4.22: Operating principle of the Sound Collector.
The Sound Collector extracts the oldest sound file recorded within the green marked time span after the keyword “sound” has been spoken.

The interface to the Sound Collector is realized as follows. Besides the obligatory object ID that allows a clear assignment to a particular discourse, the XML document (cf. below) contains the timestamp t. This timestamp is specified in milliseconds. Regarding the sampling rate (16 kHz) and the time between the feature computations within the speech recognition module (10 ms) (see S. Hohenner [Hoh05]), the resulting accuracy for the timestamp is better than 100 ms. This is accurate enough for the intended use of the Sound Collector. An additional tag ’COMPRESSTYPE’ allows to chose between different compression schemes (sound, speech) for the extracted audio file in order to improve the encoding result. Details are described in the appendix A.3 on page 124.

```
1  ...
2  <ID>
3    <ORIGIN mod="OAS">23</ORIGIN>
4  </ID>
5  <OBJECT>
6    <ID>14</ID>
7    <TIMESND>1130316024724</TIMESND>
8    <!-- COMPRESSTYPE can be either 'sound' or 'speech' -->
9    <COMPRESSTYPE>sound</COMPRESSTYPE>
10 </OBJECT>
11 ...
```

At the end of one processing cycle, no matter whether an appropriate audio file can be determined by the Sound Collector or not, the generated response of the Sound Collector contains an XML document. This document is similar to the queried XML file, except that the ’TIMESND’ tag is updated with the actually correct timestamp of the extracted audio file. Additionally, it contains the file size for easier processing within the receiving module. If the search for a corresponding sound file was successful, the file itself is included as binary data within a so-called *Composite Transport Unit* (*CTU*) as proposed by S. Wrede and col-

leagues [WFBS04]. Thus, the receiving components are able to process the audio file directly or otherwise to save it on the destination computer at an arbitrary location.

Summarizing, the sections before described the object-related representation in a visual and auditory sense. To combine this different information, an ontological XML-based structure has been developed that offers an easy and intuitive access on all relevant object information. The details of this approach are discussed next.

4.5.6 Ontological Textual Object Representation

An appropriate object format for the Long-Term Memory of the robot is a second key feature of the proposed Object Attention System. Such an object information container has to provide all relevant information that is needed to recognize a formerly learned object on the one hand, and a mechanism to store all object information that the user wants to memorize in the robot's memory on the other hand. The following exemplary XML-based object representation illustrates which content is stored and how it is structured.

```
1  <?xml version="1.0" encoding="UTF-8" ?>
2  <OBJECT>
3    <ID>10</ID>
4    <SCORE>0.78125</SCORE>
5    <TIMESTAMP>1140889113943</TIMESTAMP>
6    <BESTBEFORE>1148665102531</BESTBEFORE>
7    <TIMESTAMP_HUMAN_READ>25.2.2006, 18:38:33</TIMESTAMP_HUMAN_READ>
8    <BESTBEFORE_HUMAN_READ>26.5.2006, 19:38:22</BESTBEFORE_HUMAN_READ>
9
10   <!-- Position based on global map in [m] -->
11   <GLOBAL_POSITION>
12     <X>0</X>
13     <Y>0</Y>
14     <Z>0</Z>
15   </GLOBAL_POSITION>
16
17   <RELATIVE_POSITION>
18     <!-- (-) -> left of robot : (+) -> right of robot in [deg] -->
19     <ANGLE>0.0838469</ANGLE>
20     <HEIGHT>0.539959</HEIGHT>
21     <DISTANCE>0.96797</DISTANCE>
22   </RELATIVE_POSITION>
23
24   <FEATURES>
25     <COLOR confidence = "0.85">red</COLOR>
26     <OWNER confidence = "0.3">Axel</OWNER>
27     <SOUND>/memory/objects/10.ogg</SOUND>
28     <TYPE confidence = "0.85">laptop</TYPE>
29     <VIEW>/memory/objects/10.ppm</VIEW>
30   </FEATURES>
31   <RELATION confidence = "0.85">left to</RELATION>
32   <RELATED_TO confidence = "0.85">14</RELATED_TO>
33 </OBJECT>
```

At the beginning in line 3, the consecutively numbered *object ID* is specified. It is used to enable a common ground for the currently processed object reference. Within the Object Attention System the ID ensures a consistent data assignment over all process conditions. Furthermore, the external communication with other modules, in particular, Dialog, Gesture Recognition, Sound Collector, Scene Model, and the Modality Converter use this object ID. In the context of speech processing, the object ID can also be used for anaphoric resolution which is, however, currently not supported by the speech processing units.

Next, in line 4, a score value [0...1.0] is given which is the mathematical product of all confidence values assigned in the Short-Term Memory, see page 61. It is used to decide whether an object is already known or not. Besides a couple of features that need to match with features of already learned objects, an empirically determined value of 0.8 has proven as a reliable value for later object recognition tasks. In other words, all objects that provide a score value of a certain threshold are considered for the object recognition module.

As especially small objects, like, e.g., cups or books are most probably moved to another location from time to time and the robot is not always aware of these actions, two timestamps are included as well. In line 5, the timestamp when the object has been stored in the Long-Term Memory is specified, while the subsequent *Best Before*-timestamp limits the life cycle of the object. Although an active memory for robots is currently under development, this feature is not used yet. Nevertheless, manifold applications are imaginable, like an automatic mechanism that lets the robot forget the once stored object. This is useful to hold the memory consistent as, for instance, it usually does not make sense to store the location of easy-perishable fruits for several months. As a second application, the robot can take initiative and verify on its own, whether the object is still at its once learned location. As these timestamps are not easy to interpret for humans due to their POSIX format, the same timestamps are denoted in human-readable form as well. They have mainly been implemented for manual maintenance tasks performed by the user, but they can additionally be used to let the text-to-speech component read the dates to the user in order to inform him about upcoming update cycles.

The following block of the XML document includes the position of an object within an absolute global coordinate system of the environment. It is used for robotic platforms with navigational and localizational capabilities. For instance, the global positioning system helps to assign a unique object position even in an environment with different rooms. However, as the robotic platforms used, currently not support a positioning system, these values are set to zero in the given example. The only positions actually supported by the overall robot architecture are relative ones, related to the robot. Thus, the Object Attention System at least supports these locations as can be seen in lines 17 to 22 while the values are specified in cylindric coordinates.

The last semantic block of the textual object representation contains the learned object features, and, as far as available, references to relations related to other objects or locations, like “in front of the windows". In detail, the feature block contains all verbally specified feature types and their values, as well as the confidence values assigned by the Short-Term Memory. Furthermore, the location and the

names of learned object views and object sounds are given as well. Here, it has been proven as great advantage to use references to the actually stored data instead of an encapsulated object representation which includes textual and binary data at the same time. The advantage mainly consists in a compact data representation which improves memory queries. Additionally, the data can more easily be handled by, e.g., the object recognizer or the dialog system. However, the latter one only in cases if the learning of a view or a sound has been successfully completed before.

This XML-based object representation offers a great deal of advantages in contrast to proprietary data formats. Besides the already mentioned flexible usability, it can easily be extended and updated. Nevertheless a lot of data analysis has to be done before such a document can be generated. Thus, in the following, the realized processing strategy is, therefore, illustrated in order to point out how the algorithms are applied by the proposed Object Attention System.

4.6 Processing Strategy

In this section, the internal processing strategies of the proposed Object Attention System are explained. Therefore, the underlying control mechanism is presented first, followed by exemplary processing cycles for unknown object instances as well as for known objects. This distinction in different scene analysis strategies for unknown and known objects, respectively, is necessary as they principally differ in the image processing algorithms applied.

The overall control mechanism of the Object Attention System is realized by a *Finite State Machine (FSM)*, see Figure 4.23. It consists of seven states, while for the sake of improved clarity, the state *User Callback* is drawn several times, although it is always the same state. The directed edges that connect the states with each other are lettered with additional information about the events causing the Finite State Machine to change its state. The event *Abort* represents an exceptional event. This event is emerged by the dialog component if the user wants to abort the current interaction task. To provide a fast reaction, this event is, consequently, immediately analyzed by the Object Attention System and, hence, concerns every possible condition, no matter at what stage the processing cycle currently is. In the following this issue is, therefore, no more explicitely mentioned as it applies for all cases.

The operating principle of the Finite State Machine is subsequently described. As the distinct states mostly support a semantically separated functionality, this issue is reflected by different sections.

Autonomous Scene Exploration

After the Object Attention System has been initialized, it is in its idle state *Object Alertness*. Within this state two aspects are realized. First, an autonomous scene exploration by visual capturing the robot's vicinity. This capture process is implemented by the construction of mosaic images which will be discussed in detail later on. The capturing, however, is done only if the Person Tracking and Attention module does not detect a human in front of the robot. If a potential interaction

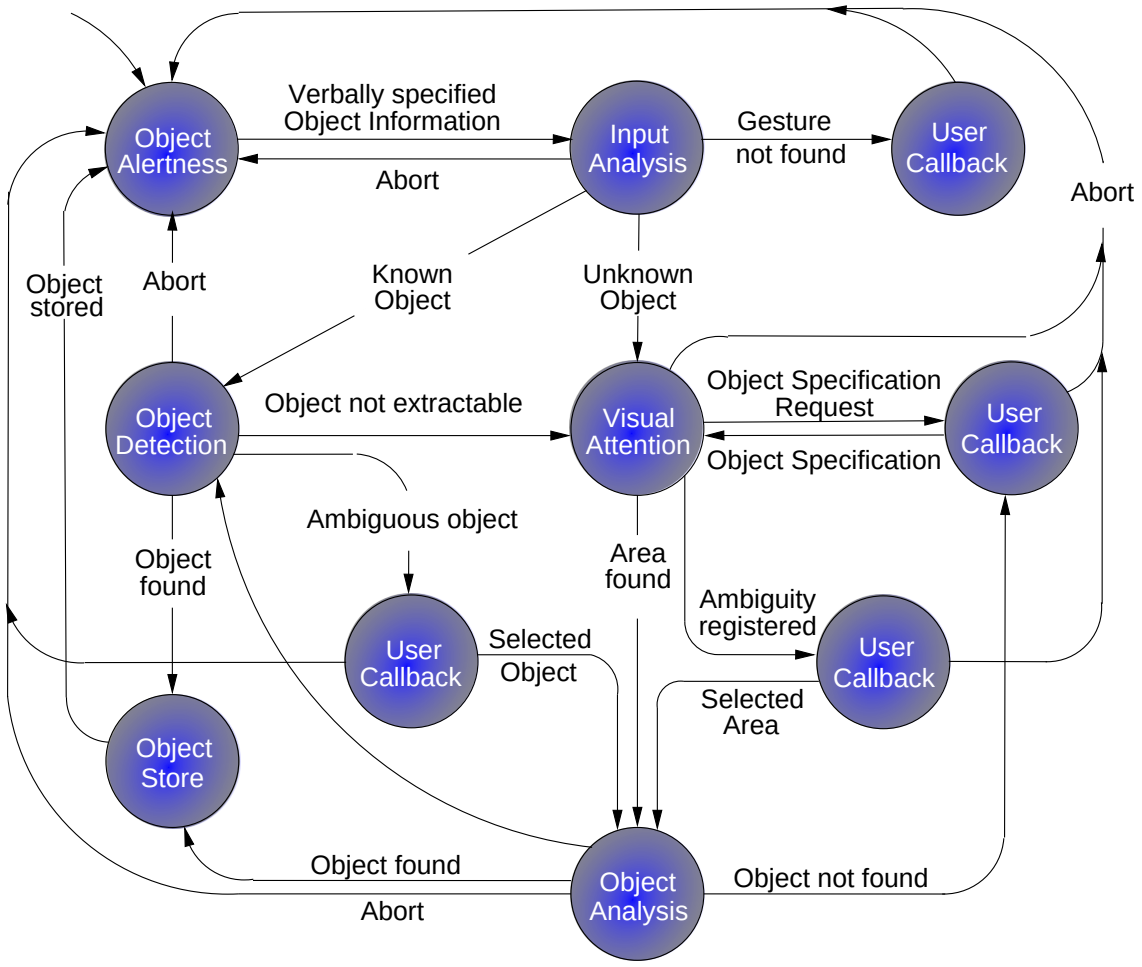


Figure 4.23: Finite State Machine of the *Object Attention System.*

partner is in its vicinity, the mosaic construction is paused until the human leaves the robot's field of view. As second task of the *Object Alertness* state, the Object Attention System continuously analyzes the input data and if new input is present, it is stored in the Short-Term Memory (cf. page 61). As soon as speech input is received, the Finite State Machine changes to the state *Input Analysis*.

Input Analysis and Temporal Modality Alignment

The input analysis of all received data begins with an interpretation of the symbolic speech data. A detailed description of the parsing routine and the processed speech data is provided on the pages 41 and 125. This enables an immediate reaction if an Abort command has been received. Otherwise, the Object Attention System continues with its normal processing routine and sends an XML document to the gesture recognition module in order to provide it with information about the current object context. In particular, the message contains three important values, the *timestamp* provided by the speech recognition system, the *object type*, e.g., cup, and the consecutively numbered *object ID* that has been assigned by the Object Attention System. After the object context has been sent, it is verified whether an *Align View* or a *Focus Object* command has been received by the Object Attention System, cf. section 3.1 on page 41.

The *Align View* command is sent by the dialog component after the user says something, like “Look here”. In particular, the command causes the Object Attention System to lower the object camera in order to show the user that the robot is ready to learn a new object and ideally awaits a pointing gesture accompanied by a verbal object description next, as it helps to resolve an object reference. Additionally, the Object Attention System returns an acknowledge message to the dialog module. This message in turn enables the generation of a verbal response for the user. In this way, the user receives an additional acoustic signal that the robot is ready to receive and interpret an new object reference. After the message has been sent, the Finite State Machine returns into the *Object Alertness* state in order to receive new messages from the dialog module.

The more complex command represents the order *Focus Object*. It is sent to the Object Attention System if the user gives a verbal object description that is optionally accompanied by a deictic gesture. Anyway, if the command is received, the camera is aligned and zoomed on the calculated position for the Region-Of-Interest as described on page 68. In the following, the symbolic information about color and size is sent to the Modality Converter and, thus, appropriately converted into numeric values. These values are then added to the *OASObject* for later use, cf. page 62 and 61. After the conversion is complete, a query process is initiated in order to verify, whether the referenced object is already known to the robot, or if the object instance needs to be newly learned.

The verification process for known objects consists of a fine-grained analysis of the textual object representations stored, described on page 80. This analysis includes pattern matching with the already stored object features *Color*, *Shape*, and *Object Type* as they are the most promising ones for a successful recognition task. If additionally the confidence value for a stored object indicates a high probability that the object can be recognized, the Object Attention System proceeds with a recognition task and, hence, its Finite State Machine changes into the state *Object Detection*. However, if no evidence is found that the object is already known, the Finite State Machine changes to the state *Visual Attention* in order to learn the object as a new instance.

Visual Scene Analysis for Unknown Object Instances

Within the state *Visual Attention* mainly the Attention Maps, e.g., for color, are calculated and a visual object view is extracted from a scene image. The particular implementation that enables this processing is very complex and has, therefore, been developed in cooperation with M. Saerbeck in his diploma thesis [Sae05]. Beginning with the calculation of the gesture-based Attention Map (cf. page 66), it is subsequently combined with the color- or depth-based Attention Map by a weighted multiplication, cf. page 69 and the following ones. As an outcome, an object view is stored for later recognition and analysis tasks. In case that the object view is not sufficiently learned, the Object Attention System can complete the view by adding further color nodes (see page 74). As this is currently not supported by the speech processing units, this is by now only possible with the Graphical User Interface, cf. page 77. In case of an error, e.g., that the color could not be detected or the robot has not yet recognized a gesture, a corresponding message is sent to the dialog component. The dialog module in turn, then, re-initiates

the object learn process as a failure discourse requires anaphoric resolution that is currently not yet supported by the speech processing units. However, if the object view has successfully been learned, the Object Attention System causes the Finite State Machine to change to the *Object Analysis* state to complete the visual learning of unknown object instances. As the recognition of already known objects is regarded as an alternative Scene Analysis step, the proceeding for a recognition task is described next, before further *Object Analysis* is discussed that applies for both cases, known and unknown object processing.

Recognizing already Known Objects

The recognition of already known objects enables on the one hand the robot to interact with objects autonomously, and on the other hand it supports interactions between the user and the robot which refer to former interactions and object information that has been stored during these interactions. This functionality is covered by the state *Object Detection*.

However, as the object recognition is not a focus aspect of the Object Attention System, a simple approach has been integrated so far to demonstrate that the proposed Object Attention System is a complete architecture that covers all relevant aspects for an Object Attention. For this reason, a separate object recognition module has been integrated, in particular, an object recognizer that is based on the *fast Normalized Cross-Correlation (fNCC)* algorithm introduced by Lewis in [Lew95]. Anyway, a lot of far more sophisticated recognizer exist, e.g., [Dic99, Bra06, SWSK05] (also cf. page 48). The basic concept behind this algorithm is a pattern matching. The learned object views are, therefore, considered as image patterns that the recognizer has to search for. As a result of a successful recognition, the recognizer visualizes the found region on the display by a squared frame. In this way, the user can directly verify that the object has been correctly recognized. After the successful recognition has been completed, the Finite State Machine proceeds with the state *Object Analysis* that allows a more detailed examination of the recognized object.

Detailed Object Analysis

An extended object analysis for visual and auditory features is performed in the state *Object Analysis*. It is used to collect as much information about an object as possible, based on its previously learned object view and an eventually existing object sound. Therefore, the Object Attention System verifies with help of the Sound Collector (cf. page 78), whether an object sound has been recorded. Thus, if a sound exists, it is stored otherwise the normal processing continues anyway. Next, the learned object view is analyzed in more detail. As the camera has been appropriately zoomed during its alignment on the calculated Region-Of-Interest, the object view usually contains detailed textural information. Therefore, a calculation of SIFT- and PCA-SIFT features [Low04, KS04] is performed. As an outcome, salient feature points are obtained that can be used for a subsequent object recognition task. As soon as the SIFT-features have been stored, the Object Attention System causes the Finite State Machine to continue with the state *Object Store*.

Long-Term Storage of Objects

The storage of object information in the Long-Term Memory is done in the state *Object Store*. Here, the XML document presented on page 86 is generated. Besides this storage task, this state is responsible for the generation of a message for the dialog component which is used to inform the user that the object learning has been successfully completed. Then, the Object Attention System returns the control for the object camera back to the Person Tracking and Attention System in order to let it track the user's face again. Last but not least, the Finite State Machine returns to the state *Object Alertness* to be able to receive new orders.

This completes the description of the overall processing cycle of the Object Attention System. As now all relevant aspects of components developed have been presented, a brief summary points out the essential features in order to lead over to issues concerning the integration in an existing robotic system.

4.7 Summary

This chapter described the most of the development aspects for the proposed Object Attention System. The chapter began with a brief discussion of related work in the field of object learning within a robotic context. Here, it has been shown that besides features, like *Color* and *Depth*, recent developments follow the paradigm of abstract object models in order to be able to recognize objects once learned. Especially graph-based approaches have been proven to work well for the description of relations either for features within a single object, but also for relations between several objects and locations. At the end of the related work section, the temporal relationship that exists between different modalities has been pointed out. As an outcome it has been shown that on the one hand a temporal dependency between deictic gestures and speech exists but on the other hand that this dependency is highly dependent on the given scenario. The related work, then, lead over to the robot hardware used, in particular, the mobile robot BIRON and the anthropomorphic robot BARTHOC. They have been described as they serve as application scenario for the Object Attention System. Thus, the available sensors for the Object Attention System have been explained as well.

The description of the components that realize the Object Attention System was the focus of the subsequent sections. At first the requirements for a flexible communication within a cognitive motivated robot were presented. It has been indicated that open standards, like XML for textual information or OpenCV for image processing tasks, are the most promising alternatives to support as much robot architectures as possible. In this context, an XML-based flexible global configuration for settings that concern the hardware setup, memory-dependent settings, and the communication links has been presented for a simple possibility to adapt the Object Attention System to new environments or robots, respectively.

The actual internal structure of the implemented biologically-inspired Short-Term Memory has been introduced as first component of the Object Attention System. It has been shown that it offers a great flexibility regarding the fusion of different modality data in order to combine them in a consistent data structure called OASObject. This included the appropriate transformation of symbolic data into

numeric values that are needed for the vision-based processing. Consequently, the concept of the Modality Converter has been introduced. The resulting values are mainly used for the determination of the Region-Of-Interest while for its localization the algorithms used to steer the robot's focus of attention have been presented as well. It has been described that a 3D-based gesture recognition provides several advantages in contrast to a 2D-based approach. Although it is currently too expensive in terms of computational power for the intended usage on a mobile robot, it has been shown that it is worth to pursue such a 3D-approach as it allows a more accurate determination of the Region-Of-Interest.

The probably most important aspect for the Object Attention System, the perception and representation of objects has been described after all interfaces and the dependencies between the different modalities have been presented. Therefore, the analysis methods for the visual object appearance have been described. One of these methods is a graph-based representation that has been developed in cooperation with M. Saerbeck. It is basically based on symbolic speech information and color-based features. Thus, a textual and visual object description became available which has been presented as well. A second vision-based approach using depth information that has been developed together with M. Köllmann has been introduced next. It has been shown that depth information is, besides *Color*, a second valuable source of information to segment an object in a given scene. In addition to the visual object representation, the audio-based object representation has been considered as well. Therefore, the principles of the Sound Collector have been introduced. It is able to add an object sound to a given textual object representation. The latter one in turn has been described as last aspect in the section about the analysis of the object perception and representation. The textual object representation is used for long-term storage and, hence, serves as part of the qualitative Scene Model. Finally, the overall processing strategies for unknown object instances and known object types, based on a Finite State Machine have been explained in detail. It has been shown how the object analysis methods were actually applied in the Object Attention System.

As the description of the actual implementation of the Object Attention System is now complete, its integration in an existing robotic environment is described in the following chapter.

5. Integration of the Object Attention System in a Robot

The integration of the proposed Object Attention System in a real robot requires an architecture that provides interfaces to sensors and other modules, like the dialog system or the gesture recognition system. Thus, the underlying architectural model of the robots BARTHOC and BIRON is discussed in this chapter. In this context, the communication framework which is responsible for the data exchange between all modules is regarded as well. Additionally, as the Object Attention System is designed to acquire as much information about objects as possible, a first interface to an iconic memory infrastructure has been realized. The approach used is based on Multi-Mosaic images which enable the acquisition of different views for one object as introduced by B. Möller in [MPH+05, Möl05]. As an outcome of the mosaic image-based scene analysis, the Object Attention System is able to store all views and other object information in a Scene Model, cf. Figure 1.3 on page 7. The Scene Model in turn serves as Long-Term Memory, storing all objects and locations that may become relevant either in Human-Robot Interactions or autonomous interaction tasks of the robot. Concluding, these topics lead to the following outline of this chapter.

This chapter begins with a short overview of related work for approaches that cope with Long-Term storage of acquired knowledge for robots. Then, a section dealing with the knowledge representation actually applied, points out the Multi-Mosaic image capturing approach by B. Möller et al. that has been connected to the Object Attention System. Subsequently, a first implementation of the applied distributed memory infrastructure is presented that matches the requirements for the realization of a qualitative Scene Model. In the second half of this chapter, the overall robot architecture used and the underlying communication framework is introduced, while the chapter closes with a brief summary.

5.1 Related Work

All cognitive motivated Personal Robots need to integrate an architecture that enables a communication between all semantically distinct units, like the speech

recognition or the visual processing components. Furthermore, such an architecture connects the different units with the available sensors and also regulates the upcoming control and data flow. In particular, four main paradigms became popular for architectural approaches as Matarić pointed out in [Mat01]. These four classes[1] are, the *deliberative* ("think hard, then act"), the *reactive* ("don't think, act"), the *hybrid* ("think and act independently in parallel"), and the *behavior-based* ("think the way you act") control. As the main application for mobile Robot Companions is defined through their interaction abilities for a Human-Robot Interaction, the purely deliberative and reactive models are inappropriate. In such a complex interaction scenario, on the one hand the robot needs to be able to think about its actions and on the other hand must be able to simultaneously react on the user's behavior. Hence, only deliberative, behavior-based or a combination of both approaches are feasible. For instance, the robot HERMES which is introduced in detail on page 28 uses a combined approach as described by Bischoff and Graefe [BG99]. Another example for a sophisticated architecture is applied in the humanoid robot ARMAR, described by Burghart and colleagues [BMS+05]. It uses a three-layered architecture that is based on a behavior-oriented model. These two examples already illustrate that the model which is finally applied, differs even it is for use in a robot. This is explainable as the field of application for the two presented robots is different. Nevertheless, they demonstrate that hybrid architectures are well-suited for a convenient Human-Robot Interaction. Because of their flexible character, such a hybrid architecture has been developed for the robots BARTHOC and BIRON by M. Kleinehagenbrock [Kle05].

Another major issue that needs to be discussed in the context of integrating the Object Attention System into a robot architecture is definitely the aspect of autonomous interactions, like fetch-and-carry-tasks. Before the robot can act on its own, it is useful to let it explore the environment first, as this significantly reduces the required amount of time for independent actions. One possibility for this is given by the generation of an environmental map like it is done by Ghidary et al. [GNS+02]. As described on page 23, their robot is able to incrementally construct a map which contains various squared image patterns that have been learned during an interaction with a user. This is one possibility to match the requirements of an autonomous task execution. However, their approach uses an indoor positioning system (similar to GPS) which reduces the robot's flexibility on a large scale. An approach that partly overcomes these limitations is presented by Kelleher and Kruijff [KK05]. They propose a model for the representation of proximity between different objects. This becomes available by the evaluation of linguistic discourse and visual object appearances that allow to produce spatial proximity expressions related to objects located in the vicinity of the robot. In this way, they are able to model even vague verbal expressions, like "at the corner" or "close by" that are often used in an interaction. To sum up, they have proposed a feasible modeling approach, but as it is restricted to single scenes, e.g., a tabletop, it is only partially suitable for the generation of a Scene Model. The question whether this approach is scalable in an even more complex environment with dozens or hundreds of objects located in several rooms remains open.

[1] Explanations have been taklen from [Mat01]

As a last example, Hois et al. [HWBR06] use a domain ontology in order to support a 3D-object recognition in combination with utterances given by a user (for details related to the object recognition, cf. page 49). Their ontology is based on a concept which is divided into several categories, in particular, *Abstracts*, *Endurants*, *Perdurants*, and *Qualities*. These categories are used to describe entities in different ways, like, e.g., the Endurants describe entities that are permanently present over time, like objects, while Perdurants describe time-related entities, like a sound event for a certain period of time. The Abstracts and Qualities finally are inhered in entities. Here, the Abstract concept expresses the value for the qualities, which in turn can contain qualities on their own. To model the ontology-based representation, they use a hierarchical structure by means of the object functions, e.g., drinking vessels, office-supply, etc. that allows a relation between similar objects, e.g., cups and bottles. Additionally, actions can be linguistically assigned, like drinking. The presented approach of Hois and colleagues is very promising but as most of it is still in the state of conceptual design, it is too early to make a final judgement. Nonetheless, their approach only supports speech and 3D-data by a laser range finder, but not a gesture recognition system or features that are easier to verbalize, e.g., color, in order to improve the convenience of a Human-Robot Interaction. Here, the proposed Object Attention System and the proposed design of the Scene Model overcomes the limitations that are given in the approach by Hois and colleagues.

Summarizing, the presented approaches show that no ideal architecture for a Human-Robot Interaction or for the representation of knowledge already exists. Although in both directions a lot of research has been conducted in the past, the chosen approaches are more or less dependent on the individual tasks. However, even though the knowledge representation is not the main focus of this thesis, the models used for the Object Attention System are presented as they provide a flexible character that allows to use them in various applications.

5.2 Knowledge Representation

For a robot that is intended to reuse the information once gathered during a Human-Robot Interaction, the acquisition of a knowledge base that acts as a Scene Model is essential. This section, therefore, describes the approaches for the incremental construction of an extensible memory framework that has been connected to the Object Attention System.

Multi-Mosaic Images as Iconic Memory

The perfect recognition of previously learned objects is still an unsolved issue, as it has been described on page 47 and the following ones, although tremendous progress has been achieved in this research field during the last 10 years. A lot of approaches are based on the visual appearance of objects. Thus, it is of great advantage if more than one view of a single object can be extracted by the robot in the course of object modeling, ideally from another perspective since objects are usually looking different from a second point of view. In order to get a second or more views of an object, it either has to be rotated or the robot has to change its current location. The latter one is often not desirable during a Human-Robot Interaction as by a movement of the robot the eye contact between the

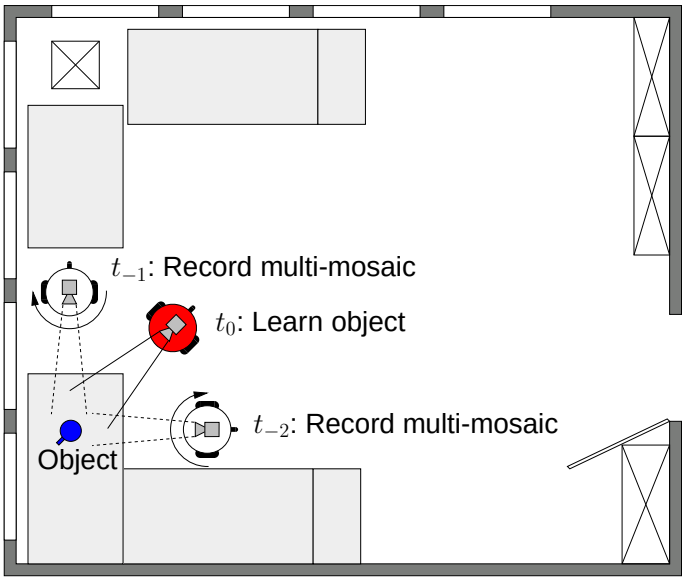
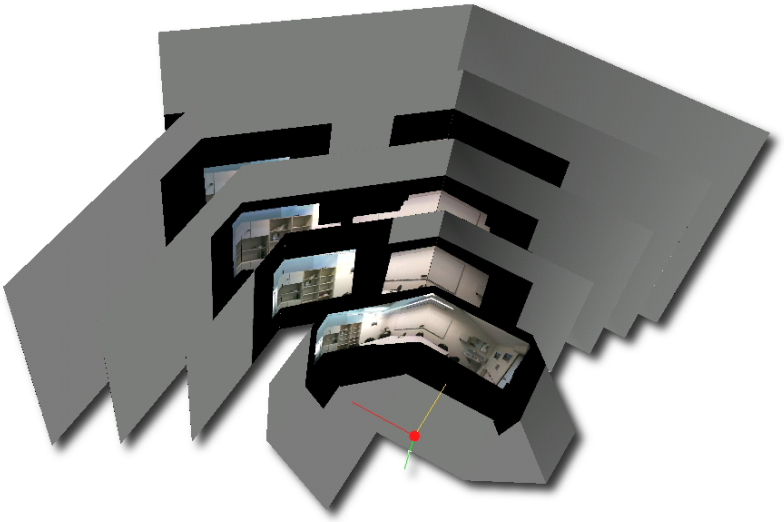


Figure 5.1: Capturing different views of objects through mosaic images.

user and the robot can break up or even worse, the robot completely loses its interaction partner. This can be avoided if the robot first scans its environment at different locations while no person is interacting with it and subsequently stores the information appropriately. The scanning during interaction pauses has got the additional advantage that usually no person occludes important parts of the scene. In this way, the robot can store a large part of a scene in an iconic memory. Thus, the robot needs to drive only to a few locations to begin the capture process as these few large images already contain all relevant locations. In an interaction at a later time then separated object views can be extracted from theses large images. Thus, the approach of Multi-Mosaic Images by B. Möller et al. is used which is responsible of the creation of the large scene images.

To do so, a corresponding examination scenario has been designed for the mobile robot BIRON which can easily be transferred to other mobile robots as well. The Figure 5.1 illustrates the basic idea. The robot moves to a distinct location at time t_{-2}. Here, it begins to capture a mosaic image as shown in Figure 5.2(a) by panning, tilting, and zooming its camera or moving its body. As soon as the mosaic image is complete, the robot moves on to a second location, e.g., another corner of a room or to another side of a table. The more different the new capturing location is with regard to the first location, the higher is the probability to be able to extract a distinct second view of an object. At this new position t_{-1}, a second mosaic image is captured. After this task is completed as well, the robot is used for an interaction with a user in the given example. During this interaction that takes place at a third location and at time t_0, another view of the object can be extracted.

While the Object Attention System passes its normal processing cycle, described in section 4.6 on page 82, the position of the Region-Of-Interest is calculated. Assuming that the robot uses a global coordinate system, this position can now be projected onto the formerly captured mosaic images at time t_{-2} and t_{-1} and an additional object view can immediately be extracted.



(a) Multi-Mosaic image with planes of different resolution.

(b) Image pattern of lowest resolution plane.

(c) Image pattern of highest resolution plane.

Figure 5.2: Illustration of Multi-Mosaic captured images.

The framework supporting this mosaicing functionality is called *Toolbox for Processing and Analyzing Images (ToPAs)* and has been developed by B. Möller and colleagues [MPH+05, Möl05]. The ToPAs framework offers interfaces that allow to extract an image patch from a given mosaic image, even in different resolutions. Examples of images with different resolutions are depicted in Figure 5.2(b), showing an image patch taken from the lowest resolution level of a captured mosaic image, while Figure 5.2(c) shows an example for an image with the highest possible resolution.

The support of different resolutions is useful as diverse applications have different requirements. For instance, a holistic localization which is currently under development for the mobile robot BIRON demands for a coarse overview of the scene. This overview should ideally be a low resolution image in order to reduce the computational costs and to minimize noise that increases with image detail. High-resolution images, however, are useful for the Object Attention System as they allow the extraction of a more detailed object view which is used for later object recognition tasks. Due to this representation of mosaic images in multiple resolutions and, hence, in multiple image planes, they are called Multi-Mosaic images. In the following, a more detailed view on the creation of such Multi-Mosaic images is presented.

The basic idea of mosaic images consists in the assumption that image sequences captured with one camera contain redundant image parts. These redundancies can be detected and used for estimating parameters that allow to warp images into a common *coordinate frame*. This procedure is called *Registration*. The underlying mathematical transformations used for the registration process are basically expressed in relation to the common coordinate frame (also called *reference frame*). Registering this image completes the *Registration phase* and enables the *Integration phase* for the mosaic image. Here, every single image is merged with the mosaic image by the fusion of color information of individual pixel. As an outcome, the mosaic image contains all fused parts of the scene that have formerly been captured as separate views by the camera and, thus, lead to a scene representation with a large field of view, see pages 82 and 91. As the mosaic images offer a larger field of view than a single captured image, distortions are unavoidable. In order to minimize these distortions, an appropriate representation needs to be selected. B. Möller chose a polytopical coordinate representation as shown in Figure 5.3. In particular, the Figure visualizes a so-called *Rhombicuboctahedron* that shows the different planes of resolutions as well in a simplified wire frame representation. This polytopical coordinate representation has mainly been chosen as it provides easier image processing than, e.g., the projection on a sphere. It goes beyond the scope of this thesis to discuss this issue in detail, however, an extensive description is given in the dissertation by B. Möller [Möl05].

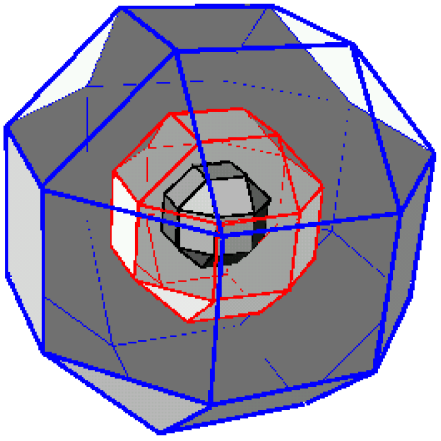


Figure 5.3: Multi-Mosaic image planes visualized as hierarchical polytopical wire frame model. The image has been taken from [MPH+05].

In brief, this plane representation has mainly been chosen as it simplifies to cope with discontinuities on the one hand, and as the planes can be regarded separately, they support the requirement of an incremental construction of the mosaic image during the *Integration phase* on the other hand. Thus, this representation allows the Object Attention System to start, stop, pause, and to resume ongoing mosaic creation without the need to discard all mosaic parts that have been fused so far. Hence, an efficient representation of the iconic data is provided.

The Long-Term representation of this iconic memory should of course be as efficient as the capturing process especially with regard to a fast and accurate data access. This is enabled by the proposed Scene Model which is described in the following section.

Scene Model as Long-Term Knowledge Base

The Scene Model offers a great diversity with regard to autonomous robot performance for, e.g., navigational tasks, and possibilities to improve a Human-Robot Interaction as it enables the robot to “remember” former interactions. The latter issue is useful for the Object Attention System as it allows to access formerly learned objects and, thus, relations can be established. But, e.g., the dialog

component can benefit from the memorized data as well, like it is described on page 78.

The basic concept of the Scene Model has been adapted from the *Visual Active Memory* developed within the EU-Project VAMPIRE [WWHB05, VAM05]. The infrastructure of the memory concept is depicted in Figure 5.4. It illustrates that the Active Memory consists of a *Memory Server* that includes the concept of *Intrinsic Memory Processes*. These intrinsic processes are used to maintain the consistency of the memory, e.g., by re-indexing or garbage collection which is comparable to forgetting obsolete data after a specified period of time. The Figure furthermore shows that the database used actually consists of two interfaces of the relational *Berkeley DB* [Ora06], one for binary data and a second for XML data which uses the native DB-XML API. Here, the coupling between the XML data and the corresponding binary data entry is realized by the *Resource Description Framework (RDF)* [W3C06].

Besides the Intrinsic Memory Processes, the access for other modules of the robot's architecture is regulated by *Extrinsic Memory Processes*. They are more loosely connected to the Memory Server than the Intrinsic Memory Processes in the context of abstractness. Thus, they can be used for higher-level tasks, like contextual or spatial reasoning. Anyway, other modules that are used in the robot's architecture are not only connected to the Active Memory but to each other as well. These connections are now discussed in more detail by the presentation of the architectural model used.

5.3 System Infrastructure

The mobile robot BIRON introduced in section 4.2 on page 56 is used for the explanations on the software architecture and its communication framework used. In particular, the interconnections between the different modules of the architecture are discussed in the following section.

Robot Architecture

The Object Attention System is part of the three-layered hybrid *System Infrastructure for Robot Companion Learning and Evolution (SIRCLE)*, depicted in Figure 5.5. This Figure provides a reduced view on the architecture as only the modules that are related to the Object Attention System are considered. At the top of the Figure, the deliberative layer is shown. This layer contains the Speech Recognition by G. Fink [Fin99], the Speech Understanding by S. Hüwel [HW06b], and the Dialog module by S. Li [LHW+05]. These modules are located on the deliberative layer as they process higher cognitive functions, e.g., the establishing of a dialog based on natural language. Thus, they are used to send orders to executing modules that are located on another layer of the architecture.

The second layer represents the intermediate level. Here, the *Execution Supervisor* by M. Kleinehagenbrock [KFS04], who designed the SIRCLE architecture [Kle05], is located as it is responsible for the communication control and data flow between most of the modules. Additionally, this layer embeds the Scene Model and the mosaic-based iconic memory which serves as knowledge base of the robot, as described in the previous section.

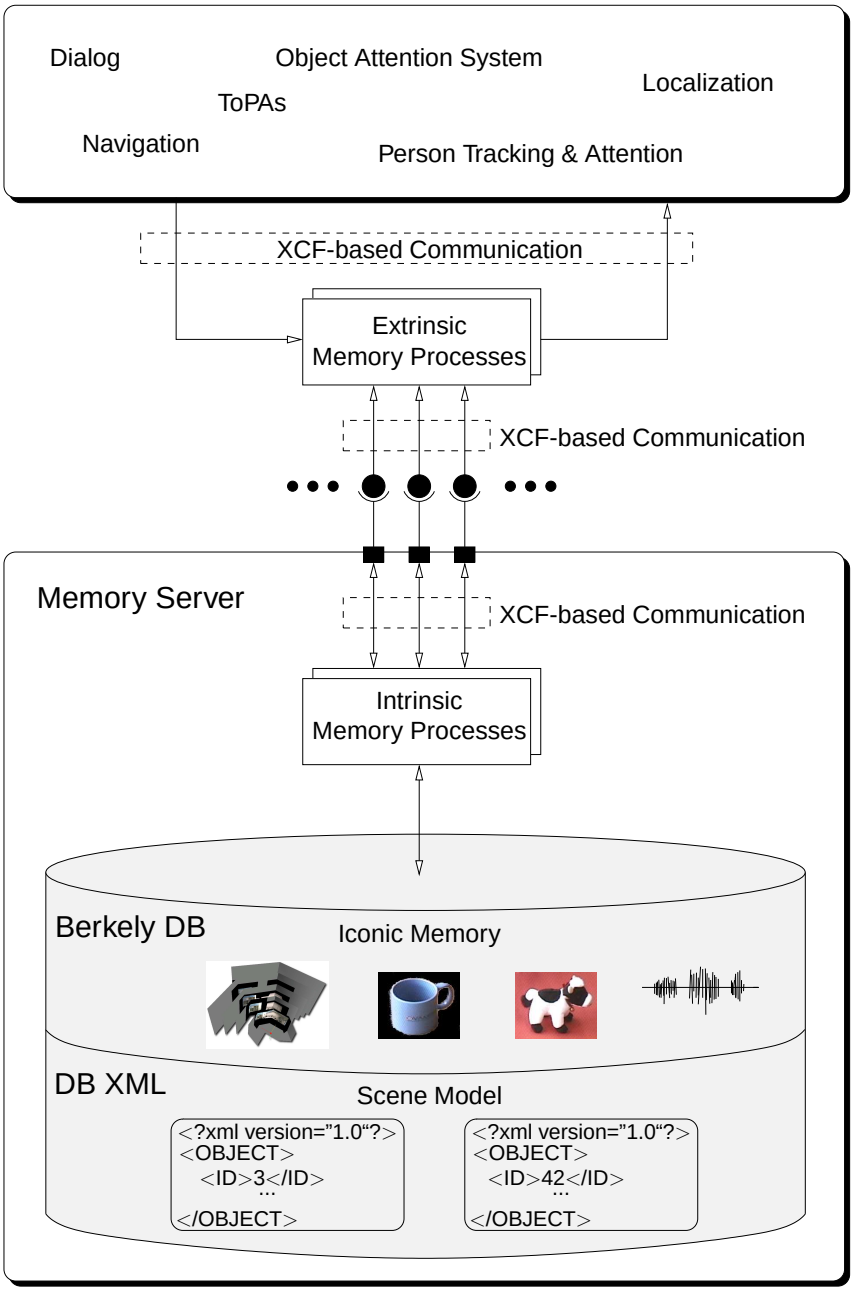


Figure 5.4: Infrastructure of the Active Memory adapted for Personal Robots.

The third and last layer is the reactive one. Here, the Object Attention System is located, while it is connected to a couple of other modules. First of all, to modules located in the intermediate layer in order to be able to communicate with the dialog module and access to the knowledge base. Secondly, to modules that are located on the reactive layer as well. In particular, these are the modules that have been described in chapter 3, 4, and 5. Additionally, the reactive layer shows the *Hardware Control* by Spexard et. al that is mainly based on the *Player/Stage software* by Gerkey, Vaughan, and Howard [GVH03, VGH03]. It provides unified interfaces for different robotic platforms and, thus, can be classified as abstraction component. It is responsible for the connection between the hardware-specific issues of the robot used (motor control, serial interfaces of the basis) and the modules that access the components connected to the robot basis. Hence, there

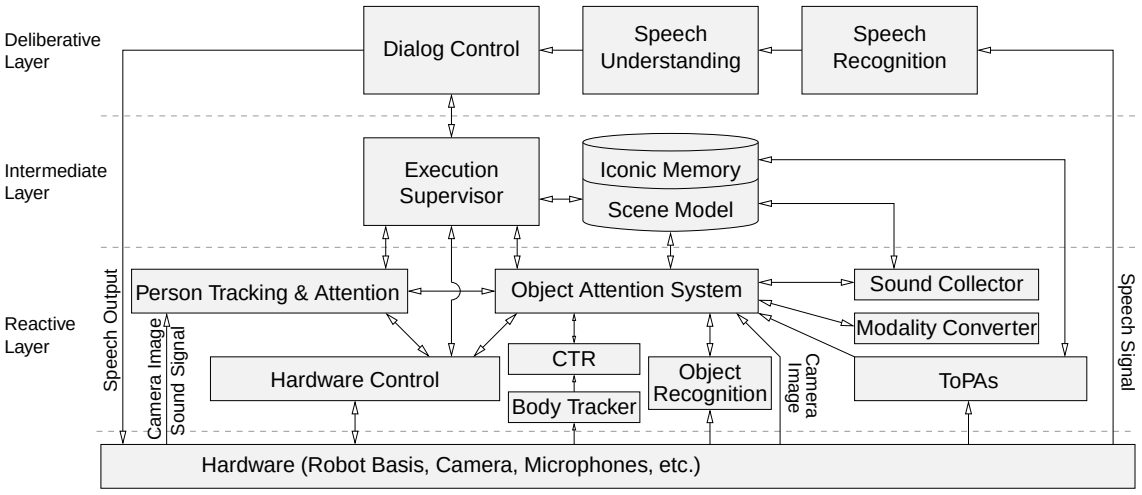


Figure 5.5: System Infrastructure for Robot Companion Learning and Evolution. The figure shows an Object Attention System-related view of SIRCLE.

is no need for the accessing modules to modify their interfaces if different robot hardware is used.

The Figure 5.5 illustrates the communication connections between the modules. The underlying communication framework is, therefore, briefly described in the subsequent section.

Communication Framework *XCF*

The proposed Object Attention System relies on the output of many other components, like the Person Tracking or the Speech Processing units. As described in chapter 3 and 4, the Object Attention System ,therefore, uses a couple of open standards for the communication, for instance, OpenCV for images or XML for textual content. Thus, it is only reasonable to use a communication framework which enables the XML-based communication per default. One approach that offers such possibilities, but also embedded binary data support represents the *XML enabled Communication Framework (XCF)* that has been mainly developed by S. Wrede and colleagues [WFBS04]. Together with the above mentioned Active Memory concept it has been introduced as part of the EU-project VAMPIRE as well. As it is from scratch designed for a cognitive motivated architecture like the SIRCLE framework, it is, hence, well-suited for the proposed Object Attention System. In the following, the concepts used by XCF are described in detail as far as they concern the Object Attention System.

Binary and XML transport

The provided XCF communication containers that are used for the Object Attention System are basically dividable into two different types. One for XML documents only and one for XML documents with attached binary data. As for all communication an XML document is used, the latter one is only for selected interfaces applied, e.g., for image and sound transport. This so-called *Composite Transport Unit (CTU)* can contain multiple binary data which are referenced by an *Uniform Resource Identifier (URI)* encoded as string. The transmission of such CTUs takes place by one of two semantic concepts provided by XCF.

Semantics for communication

In order to cover flexible module interconnectivity, the XCF framework supports two communication semantics. First, 1 *Publisher* to n *Subscriber* streams. This allows connected units to communicate with each other, while the receiving subscribers are configurable in their data-receiving behavior. In particular, this includes asynchronous and synchronous subscriber calls as well as the opportunity to fetch only the latest data set transmitted, even if more sets are buffered and not yet requested. The second communication semantic is based on *Remote Procedure Calls (RPC)* and *Remote Method Invocation (RMI)* which is realized as an n to 1 solution as well. In detail this means that various clients, so-called *Remote-Servers* are allowed to call member functions of networked *Server* processes that are provided by a single Server. Last but not least, an optional XML-Schema-based verification is included in order to ensure the validity of the data transmitted, cf. page 119.

To sum up, due to the Publisher/Subscriber and the Server/Remote-Server semantic, XCF offers easy data access and modification interfaces supported by *XPath* and *XML Query* expressions. Thus, the interfaces to other modules could be integrated in a simple manner. This completes the overview of the XML-based data handling by XCF. As last aspects concerning the communication framework, the logging mechanisms and error handling routines provided by XCF are discussed next.

System-wide Logging and Exception Handling

During the development of a software project like the Object Attention System, a lot of dependencies need to be considered, no matter whether they concern the timing between different modalities or the data format exchanged. Therefore, a fine-grained logging support is essential as many different people and even more different software modules are involved in the development of a Personal Robot. These challenges have been faced with a framework for logging and introspection support. In particular, for the logging functionality three approaches are currently supported, *Log4J* [Apa06b] for Java-based software, *Log for C++* [BGW+06] which is an early C++-port of Log4J, and *Log4cxx* [Apa06a] as successor of *Log for C++*. All three approaches allow the definition of various loggers that differ not only by their specified name, but can also be assigned to different log levels, like, e.g., *ERROR*, *WARN*, *DEBUG*, and a few more. This is especially very helpful to disable a distinct logging level in order to avoid an information overflow. A typical example of one logging message is shown below. The line begins with a timestamp in POSIX time. Then, the name of the logging level (here:INFO) is given, followed by the name of the logger. In this case, the logger's name is 'ROR.ObjCommFlow' which means that the logger belongs to the 'Resolving Object References'[2] module and is subscribed to the class 'ObjCommFlow'. The remaining part of the message contains the actual logging text.

```
1138879336 INFO ROR.ObjCommFlow : Processing time of OAS: 4258 ms
```

[2] An alternative identifier for the Object Attention System

Those logging messages can be displayed in two ways. Either directly in the command shell where the program has been started or in a system-wide and centralized logging facility, the *XCFLogger*. But, not only specific logging messages can be displayed by the XCFLogger, it is additionally possible to intercept and visualize the textual content of the data that is exchanged.

An extensive logging functionality is, however, only one important issue, especially in a large-scaled architecture for Personal Robots. At least as much important is a well-defined exception handling. This is another reason why XCF has been chosen as communication framework for the robots used as it provides a fine-grained error handling as well. Thus, the overall system becomes a great deal more robust as usually not all imaginable error cases can be considered.

5.4 Summary

In this chapter the integration of the proposed Object Attention System in an existing robot has been presented. This included at first the Long-Term representation of object knowledge. Here, the concept of Multi-Mosaic images has been pointed out and how the mosaic images are applied to get different object views. In brief, the attractiveness for additional views consists in the possibility of improved object recognition tasks that become important during subsequent tasks for the robot after it has learned referenced object instances. The second part of the knowledge representation dealed with the concept of an Active Memory which allows a rudimentary and incremental construction of a qualitative Scene Model.

After the discussion on the knowledge representation, the system infrastructure used, in particular, the robot's architecture SIRCLE and the underlying communication framework XCF have been presented. It has been shown that a hybrid robotic architecture is well-suited for the task of a natural Human-Robot Interaction. Furthermore, the main contributions of the XCF framework have been pointed out as far as they concern the Object Attention System.

Finally, after all relevant aspects for the development of the Object Attention System have been presented, an evaluation is given in the following chapter in order to prove its suitability for real robots.

6. Evaluation

In this chapter, the results of conducted evaluation experiments are presented and subsequently discussed. As the Object Attention System employs mainly deterministic approaches, like the Finite State Machine or the temporal modality fusion within the Short-Term Memory, these issues will not be discussed. It would not make any sense to roll out the reliability in additional tables as continuous code reviews during the development ensured that the Object Attention System is stable software. On many occasions has been shown that the Object Attention System is indeed operating very robust. For instance, on review scenarios for the EU-project COGNIRON, the solemnization for the Collaborative Research Center 673, and in dozens of testing sessions while more than 1000 object representations have been successfully learned. In the following, therefore, only the non-deterministic issues are regarded in detail. These issues are divided into the three categories

- Accuracy for the determination of the Region-Of-Interest
- Visual object appearance based on color features
- Depth validation of objects

These aspects reflect the technical capabilities of the Object Attention System best, as these mechanisms enable a subsequent object recognition that is, however, not considered in this evaluation chapter as it is not within the focus of this thesis. After all, this leads to the following outline of this chapter.

In the following, at first the accuracy calculations for the resolved object positions are presented in section 6.1. The subsection begins with a description of the experimental setup that served as the basis for user studies with the Object Attention System. These user experiments have been conducted in cooperation with J. Schmidt and N. Hofemann, as they needed to evaluate their systems [SKF06, Hof06] as well. The collaborative experiments offer, moreover, the opportunity to evaluate the proposed Object Attention System on authentic data

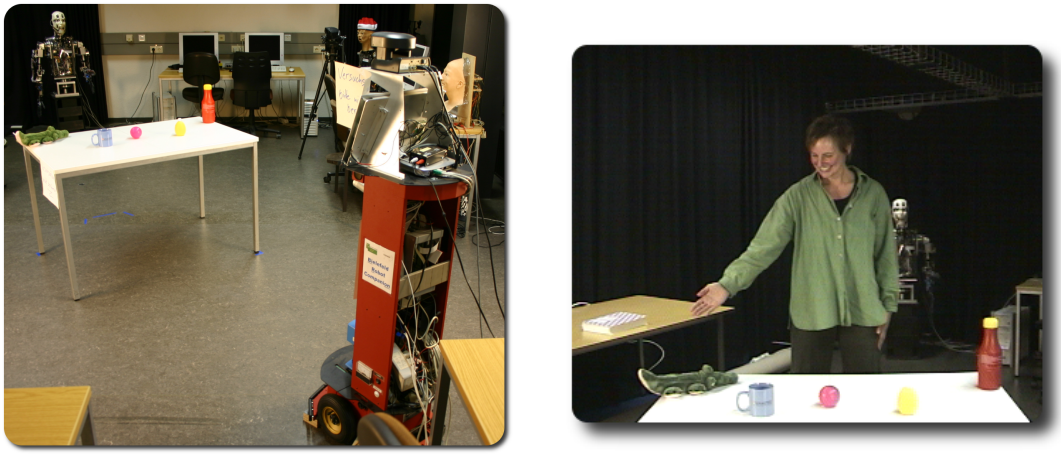
sets. After the setup in that section has been explained, the actually measured and calculated position values for the Region-Of-Interest are discussed. In particular, the values are discussed and interpreted with regard to their relevance and quality. In the subsequent sections, the results of the second and third category are presented. In detail, for the color-based object analysis, section 6.2, and the depth-based approach, section 6.3, that have been developed as diploma theses in cooperation with M. Saerbeck and M. Köllmann. At the end of this chapter, a brief summary follows which reflects the main statements of all conducted experiments.

6.1 Determination of the Region-Of-Interest

The accuracy of the determined Region-Of-Interest that is selected based on the analysis of the user's actions is the most crucial part for a successful object learning scenario. This is easy to understand because an object can only be learned if it is in the perception area of the (visual) sensors. Therefore, a lot of effort has been spent to provide a representative examination of the estimated positions for the Region-Of-Interest. The next paragraph presents the setup used for the corresponding evaluation.

Experimental Setup

To face the requirements for the 3D-Body Model Tracking System, the gesture recognition, and the Object Attention System, a specific setup with the robot BIRON has been used for the evaluation, see Figure 6.1(a).



(a) Evaluation setup for user interactions. (b) User pointing to an object.

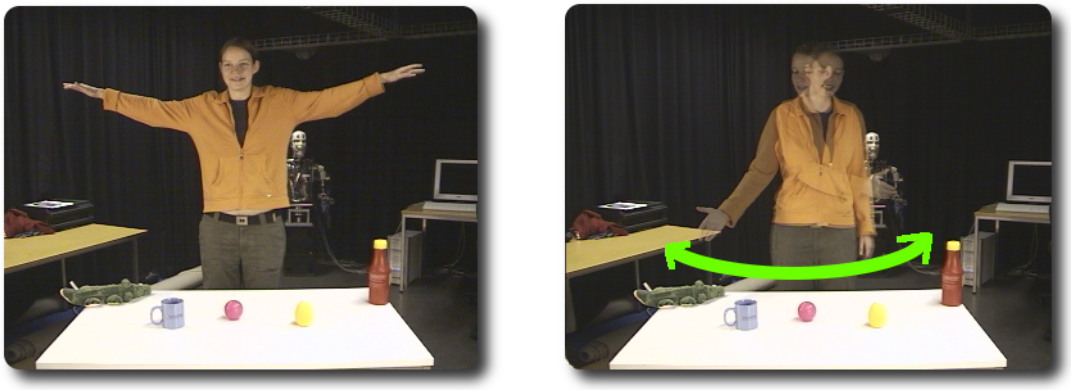
Figure 6.1: Evaluation setup for the Object Attention System
This setup has been used for the estimation of the Regions-Of-Interest.

Unfortunately, the complete evaluation needed to be performed offline for two reasons. Firstly, the available computational power was insufficient for the 3D-Body Model Tracking System to enable an online performance. Secondly, as the development of the body tracker currently focuses on tracking performance and, thus, it does not yet enable an automatic initialization for a body model. As a consequence, the positions of the different participants and objects were

manually measured in relation to the origin of the robot coordinate system, cf. page 64. For visualization of a typical pointing gesture, Figure 6.1(b) shows an exemplary gesture directed to the green toy crocodile used.

The conducted experiments were performed with 6 persons, three female and three male. Half of the participants were unexperienced and performed the gestures for the gesture recognition for the very first time. The other half were experienced users. All participants had to follow the same sequence:

1. Starfish configuration to enable initialization of the body tracker, cf. Figure 6.2(a)
2. Pointing from the bottle to the crocodile with top side of flat hand oriented to the robot (withdrawing hand after each object pointing)
3. Waving with opened right hand, inner side of hand oriented to the robot
4. Presentation of objects (from the bottle to the crocodile), cf. Figure 6.2(b)
5. Pointing from the crocodile to the bottle with inner side of flat hand oriented to the robot (withdrawing hand after each object pointing)
6. Waving with opened left hand, inner side of hand oriented to the robot
7. Presentation of objects (from the crocodile to the bottle), cf. Figure 6.2(b)
8. Repeat steps 1 to 5 three times



(a) Starfish pose for the initialization of the 3D-Body Model Tracking System.

(b) Presentation gesture (Steps 4 and 7).

Figure 6.2: Illustrative behavior for the evaluation of the ROI.

The overall sequence of the experiments has been performed as follows. The different actions were captured by the gesture camera and the object camera (cf. page 56) at first and, subsequently, stored on the hard disk. Then, the 3D-Body Model Tracking System has been manually initialized by a manual adaption of the body models used, cf. page 36. The resulting hand trajectories of the body tracker were then used as input for the gesture recognition. This enabled the training of appropriate models for the different pointing gestures. These models

were trained for each person and with three of the four runs. The fourth run was then, consequently, used for the actual gesture recognition.

All considered, the training phase for the probabilistic approaches of the 3D-Body Model Tracking and the gesture recognition as well as the manual annotation of the gesture data turned out to be very time-consuming as they took several months. However, the annotations of the videos streams are essential for the generation of appropriate body models and the following gesture recognition. Thus, the Object Attention System could at the earliest been evaluated when the recognized gestures became available. Consequently, and because the robot used for the experiments is often involved in demonstrations that make a movement of the robot and the experimental setup necessary, it was not possible to keep the same setup for the Object Attention System as it actually used during the user experiments. Therefore, the experimental setup was additionally scanned by the object camera. In particular, a program has been developed that moved the camera in the scanning area that covered all possible alignment positions for the referenced Regions-Of-Interest. The basic proceeding of the scanning process is illustrated in Figure 6.3. The detailed scans have been captured by the object camera mounted on the robot. Thus, they do not exactly match the image parts within the large image that has been taken from another point of view.

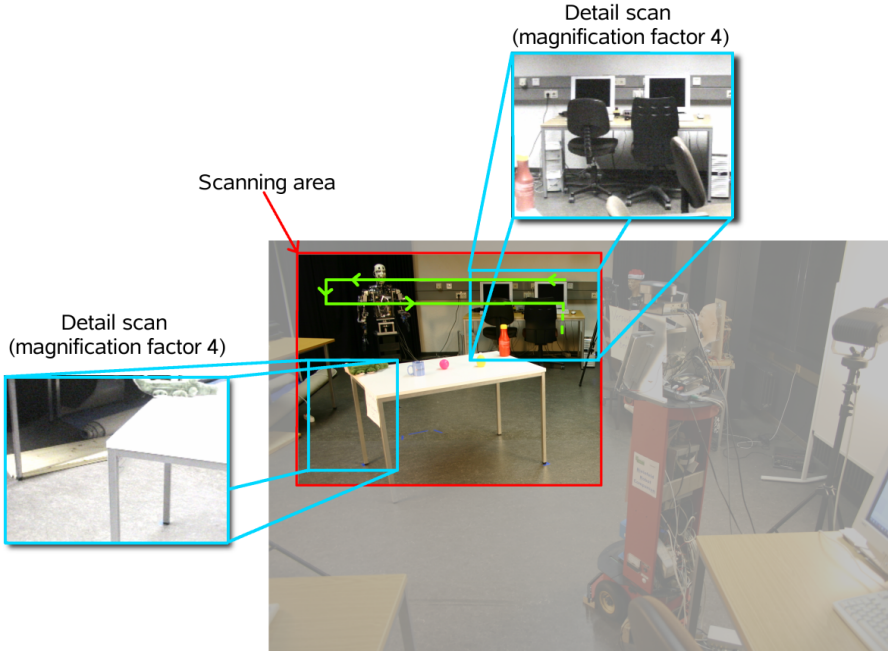


Figure 6.3: Evaluation setup scan for the Object Attention System. The green line illustrates the scanning route of the object camera.

This scanning procedure resulted in 3300 images that were stored on hard disk for later evaluation. It has to be mentioned that this scanning of course contains a lot of ambiguous positions as several locations are in the same line of sight. Nevertheless, this procedure allows an easier interpolation to verify the differently focused Regions-Of-Interest in this section. For a realistic simulation of the real interaction behavior of the Object Attention System, all images were captured with a magnification factor of 4 which allows a more detailed view of an object. This value has been chosen, as it is the default value if the user does not specify the

size which is currently always the case as the speech processing units are not able to deal with this attribute, cf. section A.2 on page 124. The following section presents the calculated results for the camera alignment and the calculated positions for the Regions-Of-Interest, respectively.

Results

During the creation of motion models for the gesture recognition it turned out that for two participants no unified model could be trained. As consequence, the Object Attention System can only process the data of the four remaining participants as for these persons individual models were trainable. The calculated positions of the Regions-Of-Interest which the users referred to, are presented in the Figures 6.4, 6.5, and 6.7. These figures visualize the mean values of all 162 successfully performed pointing gestures related to the cylindric robot coordinate system. Thus, they denote the mean values of the calculated Regions-Of-Interest as separated values for *Height*, *Distance*, and *Angle*. The objects labeled on the abscissae are ordered in the same manner as in Figure 6.1, in particular, the green crocodile with 32 gestures, the blue cup with 32 gestures, the pink ball with 30 gestures, the yellow lemon with 29 gestures, and the red bottle with 39 gestures. A detailed overview about all 576 calculated values is presented in table B.1 on page 131. In particular, it shows the 486 calculated position values for the Regions-Of-Interest while the 90 calculated values for the minimum, the mean, and the maximum error values for each object are separately presented.

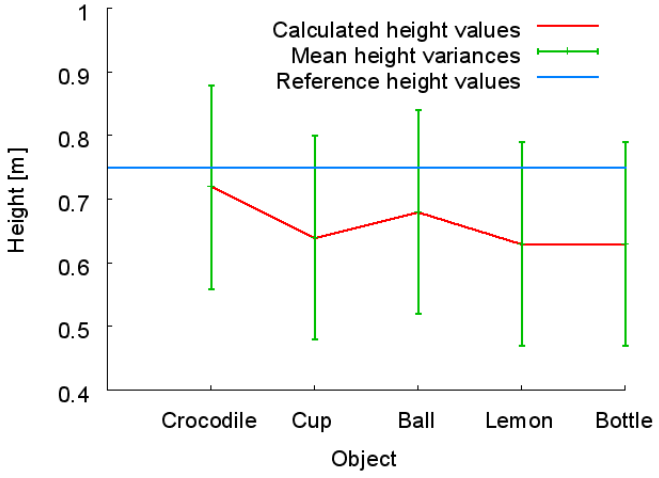


Figure 6.4: Evaluation result for the height value of the estimated Regions-Of-Interest.

Beginning with the calculated height values, the results are surprisingly accurate as Figure 6.4 illustrates. This figure, in particular, shows the calculated mean height values of all 162 successful pointing gestures. Although a few times the calculated height value is greater than the reference value (blue line at 0.75 m) the mean value evaluation shows that the mean values often are approximately 10 cm less in comparison to the actual value. These results can be explained as follows. Due to the mounting of the gesture camera that is tilted to enable the capturing of the complete upper body for the 3D-Body Model tracking framework, a measurement error for the tilt angle can be assumed. For instance, in an average distance of 2 m for the objects an error of $\pm 3^\circ$ would result in a height variation of ± 10.4 cm. This would explain the resulting error as it is very difficult to determine the exact orientation of the gesture camera due to a missing indicator on the camera mounting. Next, the calculated distance values are discussed.

The values for the distance are also often no more different than 10 cm related to the manually measured object position that, of course, contains errors as well.

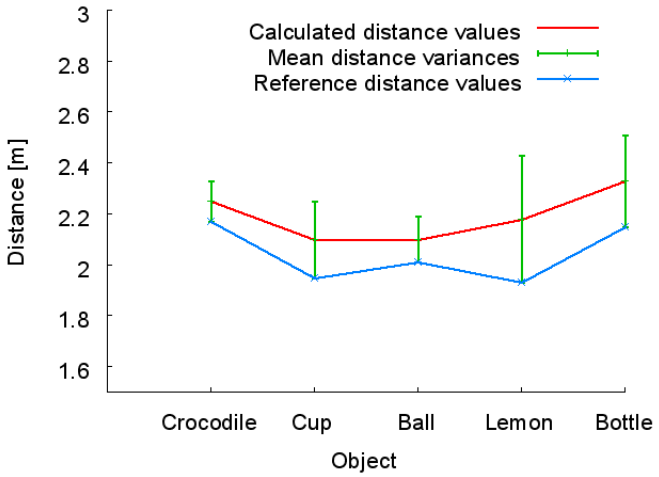
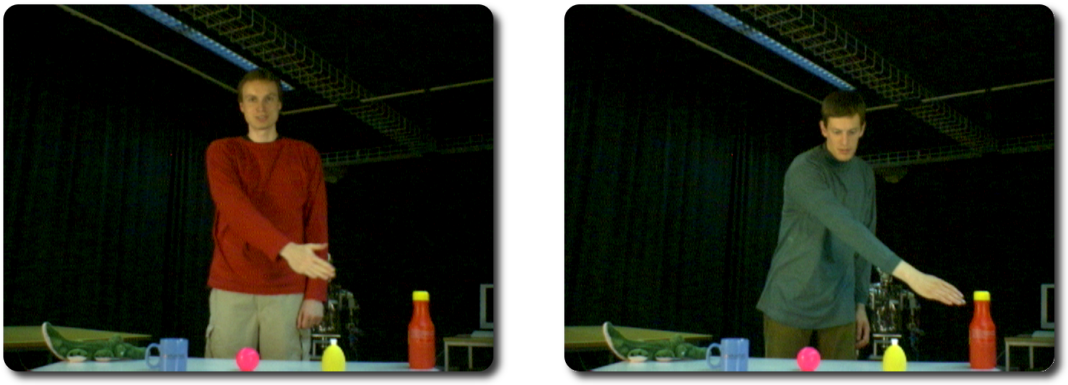


Figure 6.5: Evaluation result for the distance value of the estimated Regions-Of-Interest.

The results of the distance-related calculations are shown in Figure 6.5. The figure illustrates that the calculated values are in average 10 to 15 cm greater than the manually measured values. This can be explained as follows. Firstly, the manually measured values are related to the outer front side of the laser casing while the actual origin for the measured values is located a few centimeters within the device. Secondly, the distance of the legs slightly varied due to the pants worn by the participants. Each time, they bended forward or backward for a pointing gesture, the laser-based leg distance varied as well. Thirdly, the bending movement, of course, influenced the distance value of the participants hands, too. As a consequence, an error of $\pm$15 cm for the distance is tolerable.

Especially, the forward and backward bending of the persons explains the values that are in average 15 cm greater than the manually measured reference positions as the participants were told to directly touch the objects. This instruction has been given as a direct contact with the object probably results in the best possible values, cf. section 2.2.1 on page 18. However, the subjects often disregarded the given instruction as illustrated by the pointing gestures for the red bottle in Figure 6.6.



(a) Distant pointing gesture. (b) Near pointing gesture.

Figure 6.6: Pointing gestures with different distances between hand and object.

This last aspect of varying distances between the calculated Regions-Of-Interest and the user's hand has, additionally, a great impact on the accuracy for the cal-

culated angle values. Because the Object Attention System is not yet able to dynamically adapt to various hand ⇔ object distances, this aspect leads to the following results, depicted in Figure 6.7. The calculated values and, especially, the high variance of the red bottle shows that the user's pointing gesture indeed varied a great deal. This can be explained as all gestures were performed only with the right hand and, thus, a pointing to the most left object results in an inconvenient posture. This is supported by the constraint that the torso should be moved as less as possible in order to provide the best possible posture for the 3D-body tracker. Therefore, a tolerance value of ±30 cm can be considered as tolerable.

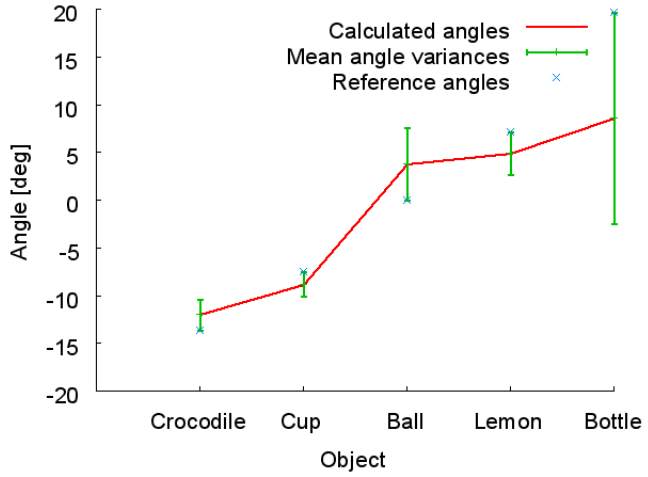


Figure 6.7: Evaluation result for the angles of the estimated Regions-Of-Interest.

However, this already large error does not explain the extreme outliers. These resulted from too fast performed gestures, especially during the presentation gesture (cf. Figure 6.2(b)) performed directly from the most left to the most right object and vice versa. As a consequence, the frames do not contain enough data for an accurate hand trajectory and a gesture recognition, respectively. As the Object Attention System considers the two frames captured before (@15 fps ≈133 ms) as well, the values suffer a horizontal shift which explains the outliers. Additionally, a last important influence causes these high errors. As the body tracker and the gesture recognition both rely on probabilistic approaches with a certain error possibility, cf. section 3.1 on page 36, these inaccuracies, finally, explain the achieved results. Although the results show in one case an error up to 87.8% (2.4° vs. 17.3° ≈ 72 cm offset@distance 2.15 m) for a calculated angle, these outliers can be classified as valid as, nevertheless, the object has been successfully focused and was fully within the camera field of view. This is attributable to the rather larger capturing field of the camera, cf. detail scans in Figure 6.3 on page 104.

Summarizing, the determination of referenced Regions-Of-Interest is accurate enough to support a convenient and robust Human-Robot Interaction and even more provides the correct Region-Of-Interest despite of inaccurate pointing positions provided by the gesture recognition modules. This is underpinned by the fact that for all performed evaluations, the referenced object was always fully within the camera's field of view while, simultaneously, a limitation on the focused area has been done. Next, the evaluation for the color-based approach is, therefore, discussed in detail.

6.2 Object Selection by their Color-based Appearance

The previous section showed that the focused area always fully contained the referenced object during the evaluation. Nonetheless, the result of a subsequent object recognition task improves if only the segmented object view without scene background is learned. This aspect has been considered in the diploma thesis of M. Saerbeck [Sae05] with whom a method for the segmentation of an object view that is based on color features has been developed, cf. section 4.5.3 on page 70. The main results of his evaluation are summarized in the following.

Experimental Setup

As experimental setup for the color-based acquisition of an object view, a selection of objects representing a variety of different colors has been analyzed. To achieve a more representative conclusion, the objects were additionally positioned in front of two backgrounds, first a white tabletop (Figure 6.8(a)) and secondly, a wooden table surface (Figure 6.8(b)). Furthermore, to emphasize the feasibility of this approach, an automatic gain control for the object camera is used to enable similar learn results under varying lighting conditions as well.

(a) Evaluation setup for object view generation with white table.

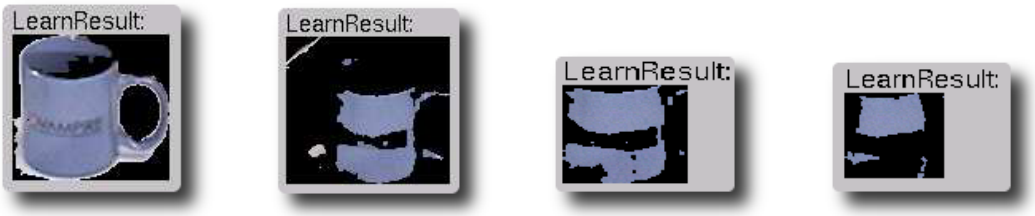
(b) Evaluation setup for object view generation with wooden table.

Figure 6.8: Evaluation setups for color-based learning of object views. The images have been taken from [Sae05].

As it would be pointless to present hundreds or even thousands of objects with a different color each, and variations in lighting, here, only two objects are exemplary analyzed to prove that the principal concept is feasible. The first object is the blue cup and the second object is the yellow lemon, as shown in Figure 6.8. The experiments conducted on these objects are, in particular, the influence of the variance value and the range of tolerable hue variations.

Results

An object view that is extracted simply based on the color appearance is often incomplete. In Figure 6.9 a learned object view illustrates the influence of the selected hue value in relation to the completeness of the object.



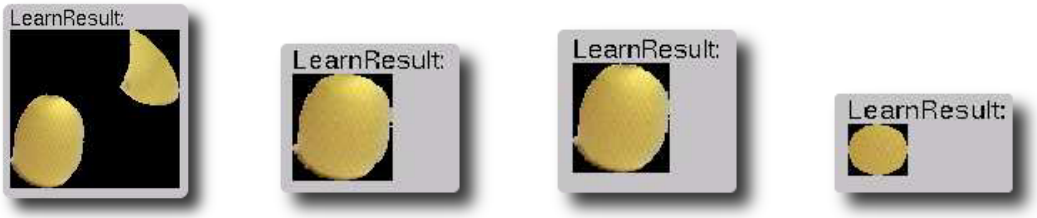
(a) Object view learned with hue value 160.
(b) Object view learned with hue value 255.
(c) Object view learned with hue value 68.
(d) Object view learned with hue value 55.

Figure 6.9: Influence of the hue value for a learned object view.
The images have been taken from [Sae05].

As the images show, does the hue value have a great impact on the completeness of an object view. The particular values used for this example are given in table 6.2. It shows that even a misaligned value that differs more than 35% from the optimal hue value can produce object view that contain more than $\frac{2}{3}$ of the object. This is often sufficient for an object recognizer to successfully select the searched object.

	(a)	(b)	(c)	(d)
Hue	160	255	68	55
Difference	0	+95	-92	-105
Difference [%]	0	+37.3	-36.1	-41.2

The second experiment investigates the standard deviation σ (cf. page 66) of the Gaussian distribution that is layed over the Region-Of-Interest to reduce the influence of pixel that probably do not belong to the regarded object as they are located in a certain distance to the center of the Region-Of-Interest. The Figure 6.10 illustrates the impact of the chosen variance on the possibility to support a resolution of ambiguities. Here, the values σ = {70, 40, 25, 15} have been used. As intended, the parameter σ is well-suited to support the feature *Size* of an object.



(a) Learned object view with σ=70.
(b) Learned object view with σ=40.
(c) Learned object view with σ=25.
(d) Learned object view with σ=15.

Figure 6.10: Learned object view of a lemon.
The images illustrate the influence of the Gaussian distribution-related variance and have been taken from [Sae05].

As an overall summary for the color-based object view generation two main aspects have to be mentioned. First, both concepts, the use of the HSI-color space

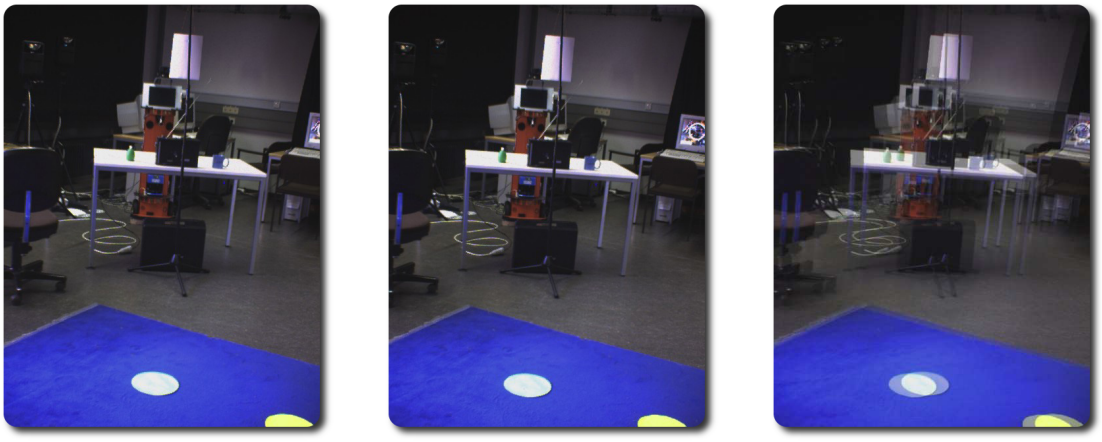
and the Gaussian distribution for the size determination of an object have shown that they support good results. However, the evaluation took place only in a laboratory environment and, thus, may cause very different results in other domains. Hence, especially the hue-value optimization needs to be further investigated. Next, the second, depth-based object segmentation approach is described.

6.3 Qualitative Measurement of Object-related Depth Values

The quality of the calculated depth-based Attention Map (cf. page 72) has been evaluated in cooperation with M. Köllmann in his diploma thesis [Köl06] while the most important results are summarized in this section. To cover the accuracy aspects for the stereo-based measurement of object positions, two different evaluations are subsequently discussed that are both related to experiences with the anthropomorphic robot BARTHOC, cf. page 58.

Experimental Setup

All measured sensor values offer a tolerance due to mechanical inaccuracies. Therefore, M. Köllmann has determined the maximal possible depth variation for the camera sensors of BARTHOC. Figure 6.11 illustrates the underlying principles. In image 6.11(a), the left camera has been aligned to its straightforward position (sensor '0'-position) at time t_0. In this case, the camera has been aligned by rotation from its, e.g., most left alignment position to the '0'-position in clockwise direction. The second image 6.11(b) was captured with the same sensor '0'-position at time t_1 but in this case, the camera has been aligned from the opposite camera orientation related to the first alignment. For the given example, by rotation from the most right alignment position to the '0'-position in counterclockwise direction. In this way, the two images can be overlayed with each other, Figure 6.11(c). The image clearly shows the difference between the two images that occurs when the camera is aligned to its initial position from opposite directions.



(a) Left camera, aligned to straightforward pos. at time t_0.

(b) Left camera, aligned to straightforward pos. at time t_1.

(c) Overlay of Figure 6.11(a) and Figure 6.11(b).

Figure 6.11: Accuracy determination of the camera sensors used within the anthropomorphic robot BARTHOC. These images have been taken from [Köl06].

This inaccuracy demands for a detailed analysis of the calculated depth values for different objects. Thus, a second experiment with a couple of objects has been conducted. In detail, the objects are a white cup, a black cup, a yellow lemon, a red bottle, a black remote control, a clock, a blue mould, a green candle, a red apple, and a red cherry as shown in Figure 6.12. The various colors and sizes of the objects demonstrate that the approach not only operates on a single object under optimized conditions but that the approach is feasible within limits for a natural environment, like a private home with various different objects. The most important limit concerns textureless scenarios. In such scenarios, the determination of object positions will not properly work as the stereo correspondence is based on feature points that can only be calculated on salient areas.

Figure 6.12: Objects serving for the evaluation of the depth-based *Attention Map* and the acquisition of object positions. The images have been taken from [Köl06].

The following section summarizes the results of the experiments above described for the sensor accuracy determination and the acquisition of object positions.

Results

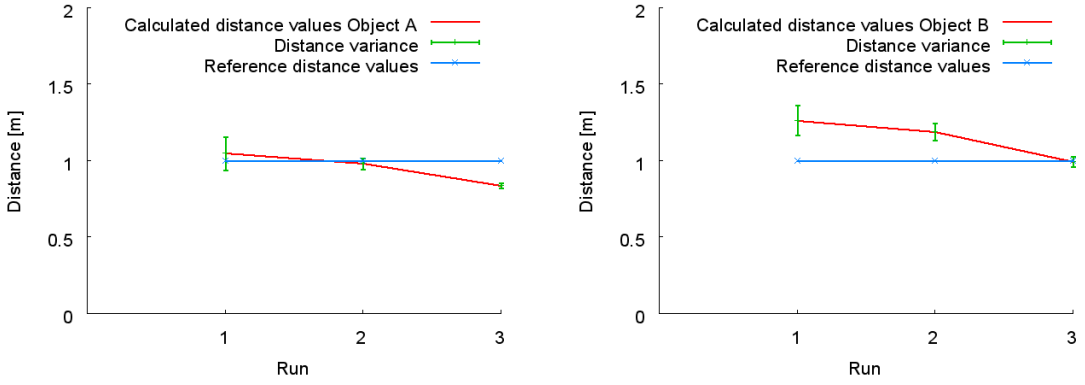
The analysis on pixel level of Figure 6.11(a) in relation to Figure 6.11(b) showed a shift of U_x = 22 pixel for two corresponding pixel. This results in an inaccuracy α_U for a given camera calibration matrix K_l in an angle of 0.017 rad ($\approx$ 0.974°) as the following equation 6.1 illustrates. The variable f_x denotes the magnification in the horizontal x-direction.

$$\alpha_U = arctan\left(\frac{U_x}{f_x}\right) \approx 0.017 \tag{6.1}$$

To point out the influence of this inaccuracy, the experiment refers to an object in a distance of 1 m. While the stereo basis of the left and the right camera is 60 mm, a corresponding point on the object results in a panned camera orientation of 0.03 rad ($\approx$ 1.719°) for each camera. This leads to the following equation:

$$0,03 - 0,017 = 0,013 \quad 0,03 + 0,017 = 0,047 \tag{6.2}$$

A triangulation of the values 0.013 and 0.047 results in maximal possible distances of 0.64 m and 2.31 m for the object located in a distance of 1 m. This example demonstrates that the sensor accuracy provided by the robot BARTHOC might be not sufficient for the extraction of a depth-based object appearance. Therefore, the second experiment (cf. Figure 6.12) has been conducted in order to get insights on the influence of the sensor accuracy for different real objects. To reduce the effect of outliers, three runs have been evaluated for the 10 objects that have always been arranged in pairs. Additionally, the positions of the two objects have been swapped (Object A → Object B and vice versa). In this way, an overall amount of $10 \cdot 3 \cdot 2 = 60$ object positions has been achieved which is, in detail, shown in table B.2 on page 134. For a clearer representation, however, the results are visualized in Figure 6.13 and Figure 6.14.



(a) Results of depth values for object A.

(b) Results of depth values for object B.

Figure 6.13: Evaluation results of depth calculations for two objects.

The diagrams clearly show that the influence of the inaccurate sensors is acceptable as the depth variation of all calculated object distances is less than 11 cm. Consequently, the result is even better than the probabilistic depth estimation of the 3D-Body Model Tracking System which uses one monocular camera only. In addition to the distance evaluation concerning the object position in relation to the camera, the distance between the two objects has been investigated as well. For a reliable statement, this distance of 30 cm has been manually measured. The Figure 6.14 visualizes the achieved results.

It can be seen that the variance is less than 11 cm as well. However, if the first run is eliminated the achieved results are a great deal more accurate with, like the

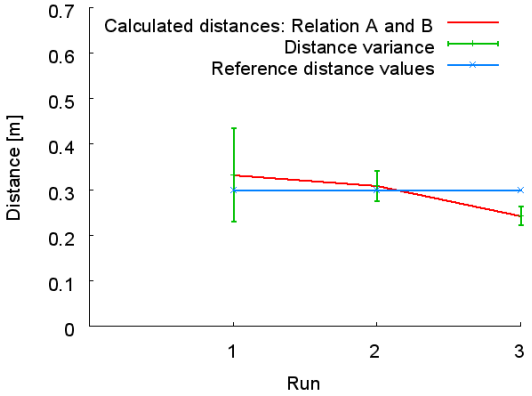


Figure 6.14: Experimental evaluation concerning the relative distance values between the object A and the object B that have been investigated simultaneously.

variance indicates with a value of less than 4 cm during the second and the third run, cf. table B.2 on page 134. These results lead to the following conclusion.

Summarizing the results, the evaluation has shown that the second and the third of all three runs are significantly more accurate than the first run. Although this behavior was reproducible during the evaluation, the reason for this could not be clarified. Most probably, the drive belts of the stepper motors produce a slip.

6.4 Summary

In this chapter, the feasibility of the Object Attention System was investigated from different point of views. In particular, all included non-deterministic approaches used have been evaluated. This includes first of all the determination of an appropriate Region-Of-Interest based on the results of the gesture recognition input. As expected, the accuracy of the finally determined region is much influenced by the quality of the gesture recognition. Nevertheless, even with provided locations occurring in a single gesture, the impact could successfully be reduced by the consideration of currently 3 time steps. In this way, the referenced object was in all cases fully within the field of view of the object camera.

The second part of the evaluation contains experiments of the color-based learning of object views. The very brief discussion on that aspect already showed that several weaknesses can be compensated as long as stable environmental conditions are provided. Especially the influence of varying lighting conditions demands for further investigations although a color model that separates the hue value and the saturation/intensity values has been used. Nevertheless is the choice to use a color-based approach unavoidable as the color is a feature easy to communicate for the user of a robot. Due to the often not sufficient extraction of the object contour, the third part of the evaluation contains the quality results of the depth-based approach used.

For the use of anthropomorphic robots, the choice to estimate the position of an object and additionally to extract the object's contour based on an eye-like stereo camera is a logical conclusion. In this thesis, the approach used has shown that it works fine even under the given preconditions of inaccurate sensors that are used for the camera alignment. Nonetheless have the results shown that the approach is more usable to estimate an object's position than its contour.

With the discussion of the previously acquired results this chapter and the description of the proposed Object Attention System is closed. In the following, an overall summary and an outlook for future work is given in order to point out the most essential statements made in this thesis.

7. Summary and Outlook

Summary

The goal of this thesis was to present a novel and flexible approach to realize Object Attention in mobile robots in order to be able to acquire a qualitative Scene Model with newly learned objects and locations.

The aim of this work has been motivated by the unstoppable development of an aging society. The trend that more and more old people (have to) decide to live alone at their homes due to an emerging lack of nursing staff seems unavoidable mainly because of the regressive birth rate. As a consequence, new solutions are needed. For instance, the development of a companion-like Personal Robot that is able to support people in, e.g., domestic domains. These domains, however, cannot be learned in advance by the robot due to their dynamic and cluttered character which demands for a learning model of Object Attention. In order to enable the best possible realization of such an Object Attention mechanism an extensive literature research has pointed out that the proposed Object Attention System requires multimodal input support. Furthermore, the robot has to be able to show its reaction to the user in order to give positive or negative feedback for the focused object or location that has been referenced. Thus, it becomes necessary to establish a Shared Attention, like it is common in Human-Human Interactions. Especially, the support for naturally spoken speech and gestures is of great relevance as a Personal Robot with Object Attention capabilities has to behave as intuitive as possible to be accepted in the human society. However, not all modalities are preferred in the same manner as research showed.

This leads to the conclusion that the Object Attention System has to support at least the most preferred modalities. But at the same time the system needs to be modular enough to enable an easy extension for additional modalities, for instance, to cover even special fields of applications that eventually require, e.g., tactile support. For this reason, the proposed Object Attention System supports *Speech, Gestures, touch-sensitive displays*, and *written commands*. But, as the object information provided be these modalities might not be accurate enough or

it is not even specified at all, the Object Attention System complements the information by an optional analysis of the currently interacting Person-Of-Interest. This analysis contains, in particular, a person localization that allows to infer the object's position and, additionally, a face identification to assign the object's owner.

The modular paradigm of the proposed Object Attention System is stressed by an unified interface that is easy to extend and to configure through the use of the extensible markup language (XML). The benefits are obvious as XML-based communication is already a widespread application in information processing. It has been shown how the XML-based communication with and within the Object Attention System is combined with the acquisition of expressive auditory and visual object information. For this object information it is essential that it is easy to verbalize as spoken language is the preferred interaction modality. Hence, the features *Color, Size, Shape*, and *Relation* are considered for the learning of a priori unknown object instances as well as for the recognition of already known objects. As an outcome, a fused textual ontology has been presented as a unit that allows the construction of a qualitative Scene Model.

The Scene Model in turn is based on an Active Memory approach that is capable to maintain data consistency on its own by continuous intrinsic data processing. Besides, to extend the reliability of autonomous interactions by the mobile robot, a Multi-Mosaic approach has been described and how it is connected to the Object Attention System. The mosaic images offer a greater field of view on the scene at a particular position. Assuming that mosaic images have previously been captured at different locations, this proceeding enables the acquisition of additional object views which can be used for a more robust object recognition.

In order to prove the feasibility of the proposed Object Attention System, it has been integrated in a robotic architecture that is used in two different robot platforms. Firstly, the anthropomorphic robot BARTHOC and, secondly, the mobile robot BIRON. Consequently, the interfaces of the Object Attention System to other modules, like the Gesture Recognition or Speech Processing have been summarized by a brief introduction to the communication framework XCF. At the end of this thesis, the performance of the proposed Object Attention System has been evaluated in detail.

The evaluation results have shown that the combination of the chosen modalities is excellent for the learning of unknown objects. This statement could be confirmed by the achieved experimental results that have been discussed from different functional perspectives. Firstly, an accuracy determination of the selected Region-Of-Interest that has been referenced by the user, secondly, an evaluation of the color-based attention processing and, thirdly, an evaluation of the stereo-based object position acquisition. For the determination of the Region-Of-Interest, an elaborate user study has been performed where 162 gestures have been analyzed for the Object Attention System. In all cases the alignment of the object camera in order to focus the referenced Region-Of-Interest was successful although the referenced object was not always in the center of the field of view. Nevertheless, this alignment can be regarded as sufficient, as the objects were in each case fully within the camera's perceptional area and, thus, an object view could be learned. However, the variance was mainly caused in consequence of inaccurate pointing gestures and uncertainties in the gesture processing modules

that apply probabilistic models. The second part of the evaluation dealed with the calculation of a color-based Attention Map. It has been shown that in constraint environments (constant lighting conditions, bright colors) the underlying algorithm works fine. Last but not least, the stereo-based determination of object positions has been investigated. The position estimation implemented is able to deal with stereo cameras that consist of two cameras which have no fixed stereo base and can be rotated in relation to each other. Thus, the approach is predestinated for anthropomorphic setups as human eyes are separately movable within limits as well. The conducted experiments showed that even though the sensors offer a large tolerance in terms of positioning accuracy, the acquired object positions are nonetheless acceptable.

To sum up, the proposed Object Attention System provides a variety of interaction interfaces that allow a convenient and effective Human-Robot Interaction. The achieved results clearly show that the chosen modalities in combination with a flexible XML-based communication and the evaluation of the object attributes *Color, Size, Shape*, and *Relation* are well-suited for task-oriented Object Attention in interacting robots. By its flexible communication interfaces and modular software design the proposed Object Attention System is easy to adapt to different robots, like it has been demonstrated for the robots BARTHOC and BIRON. Consequently, the proposed Object Attention System provides a solid base for future work on intuitive Human-Robot interfaces.

Outlook

The proposed *Object Attention System* offers a lot of possibilities for extensions. This is especially supported by its modular implementation and due to consistent use of the *XML*-based communication format. For instance, as depth often provides an excellent source for the shape detection of objects, the visual shape analysis of an object might improve the performance of the Object Attention System as, currently, the attribute *Shape* is used only in terms of a semantic speech analysis.

A second improvement would be the integration of an adaptive mechanism that allows varying distances between the user's hand and the referenced object. This could be realized, in particular, based on accelerations of the hand as coarse pointing gestures directed to vague locations or large objects (e.g., a fridge) are often performed faster than a precise pointing gesture on a strictly localized position (e.g., pointing on a screw).

A further performance improvement of the Object Attention System could be achieved by the use of fuzzy expressions for the color-based object segmentation. Utterances, like "a little bit brighter" or "a few centimeters to the right" are not covered yet. For complex scenes this, however, might simplify the interaction without the need of using the touch screen. In this way, the Object Attention System allows an even more intuitive Human-Robot Interaction.

A. Details on Implementation

In this chapter, a more detailed and technical description is presented for the different communication formats used, the Modality Converter, the Sound Collector, and the Short-Term Memory. In this way, the internal processing of these modules can be discussed without a loss of clarity in chapter 4 that mediates the basics of the proposed Object Attention System. In some places the following explanations might be redundant with regard to the content of chapter 4, however, this is unavoidable for a clear discussion.

A.1 Flexible Communication Formats

The next section contains a detailed discussion on the different communication formats that are preferred for the use within the Object Attention System. In particular, the following description covers the aspects of auditory, textual, and visual communication issues.

Audio-based Communication

For audio-based content, the Object Attention System uses the Ogg [Fou06] multimedia container format. It is non-licensed, patent-free, and supported by the most important operating systems. In particular, the Ogg format is used within the Sound Collector that is responsible for the extraction of object sound, cf. section 4.5.5 on page 78. Besides the audio signal of an object, e.g., symbolic speech data is exchanged as well, cf. page 41. Therefore, the next paragraph gives a brief discussion on the format used for textual data representation.

Text-based Communication

An approach with a high rate for textual data exchange in the robotic community is provided by markup languages which belong to the *SGML (Standard Generalized Markup Language)*. Therefore, most people consider that the advantages of SGML-based approaches outweigh its disadvantages. In particular, it is a combination of several factors which makes *SGML* or a subset of it the best choice. Some of these advantages are independency of the operating system used, the

representation in a simultaneously human- and machine-readable format, or that it consists of a logically-verifiable format. A very popular subset represents the *eXtensible Markup Language (XML)* [Wor06a] that has become widely accepted and used in distributed networks, like the Internet. Although it is not patent-free, it has become an open standard which in turn is based on several internationally accepted standards, like, e.g., *XML Query* or *XML Schema*.

From this perspective, the proposed Object Attention System uses XML for three different kinds of applications

- Communication format for inter-module data exchange
- Specification of learned object files
- Specification of configuration files

As the discussion of the concrete specification for objects and configuration files has been presented in chapter 4.5.6 on page 80 and the following ones, this section focuses on the communication aspects of the Object Attention System only. Starting from the communication as it has been introduced in chapter 3 on page 33, several XML fragments of exchanged XML documents for inter-module communication have already been presented. All of them are embedded in a well-defined format in order to enable the possibility to verify the included information for validity. This is done with the specification language XML Schema as the following excerpt of a communication document illustrates.

```
1 <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
2 <MSG xmlns:xs="http://www.w3.org/2001/XMLSchema-instance" \
3       xs:type="data type name">
4   <GENERATOR>OAS</GENERATOR>
5   <TIMESTAMP>1160579023626</TIMESTAMP>
6   ...
7 </MSG>
```

After the initial declaration of an XML document, the actual content begins in line 2 with the tag 'MSG'. It encapsulates all exchanged data. Furthermore, this tag contains some specifications, in particular, the namespace used as it is defined by the attribute 'xmlns:xs'. Besides, at the end of the line, the related data type is specified as well (cf. [Wor06b] for details). Next, in line 4, the 'GENERATOR' tag is specified which contains the name of the sending module. In line 5, a globally defined timestamp denotes the time when the data has left the sending module in POSIX time. This timestamp entry marks the end of the general frame description of exchanged data. Subsequently, the document contains the individual data. This is symbolically indicated in line 6 before the frame declaration ends with a closing 'MSG' tag.

One tremendous advantage of this XML-based approach is that it enables the use of arbitrary languages by an encoding in unicode format. For instance, Asian fonts which are not definable within the 8 bit ASCII standard can be integrated in this way, too.

Vision-based Communication

Regarding the vision-based processing, a distinction between separate images and image streams or videos streams has to be made. Since both categories are used by the Object Attention System, the formats used are briefly discussed in the following.

For capturing the robot's environment, the Object Attention System uses the pan-tilt camera described in section 4.2 on page 57. While the camera produces a video stream encoded in the YUV-Format, the stream is converted by a plugin for the graphical plugin shell *iceWing*, that has been developed by F. Lömker [LWHF06, Löm06], into a sequence of OpenCV-based *image processing library (IPL)* image format images [int06]. The OpenCV library by intel has been chosen, as it is on the one hand freely available for the most common operating systems, Linux, MacOS X, and Windows and on the other hand because it supports a broad selection of certain image formats and algorithms. Within the Object Attention System, the IPL images are used to access and to modify the image content.

At the end of the image processing different object views need to be stored as part of the object representation for the Long-Term Memory. Here, the *Portable Pixel Map (PPM)* format is used for image representations that are stored for later interactions. The reason for this choice is that it represents a lowest common denominator color image format for single image files and, thus, is easy to analyze and to construct.

Summarizing, the chosen open standards for the different modality data enable a well-documented integration of the Object Attention System in other robotic systems.

A.2 Modality Converter

The Modality Converter supports four different features (Color, Relation, Shape, and Size) that can be transformed for the use of visual object processing within the Object Attention System, cf. page 62. The following sections provide a detailed insight into the XML-based notation from a semantic point of view.

Color

The predicate *Color* is the most important symbolic information for the object learning algorithm of the Object Attention System. Therefore, it offers the most flexibility of all considered predicates. The listing below illustrates the entries used for the symbolic $\Leftrightarrow$ numeric conversion. As shown in lines 2 and 14, all entries are embedded in a *COLOR* tag within the XML-based lookup table. Within the color tags in turn, each color is specified within a *SYMBOLIC* tag which has an additional attribute *name* that contains the symbolic description of the speech processing unit. This name is evaluated by an *XPath* expression in case of a query. In line 4, the name of the color space/model used is specified. The following lines then contain the actual numeric values, while each color space channel, e.g., the first one (here: A) has got a minimum and a maximum value. This becomes necessary, as the color perception of humans and the robot varies, e.g., due to varying lighting conditions. For details, please see the dissertation of Backer [Bac04]

who did an extensive analysis of this aspect. With help of these minimum/maximum values the recognition results can be significantly improved. As last entry, a *CHANNELS* tag is included which is used to select the channels (lines 5 to 10) to be considered for computations. In particular, the channels value is binary encoded (8-4-2-1 code). For instance, if only channel A should be considered, then the channels value is 1. A combination of all three channels would, therefore, result in the value 8.

```
1  ...
2  <COLOR>
3    <SYMBOLIC name="red">
4      <MODEL name="HSV">
5        <VALUE_A_MIN>0</VALUE_A_MIN>
6        <VALUE_A_MAX>4</VALUE_A_MAX>
7        <VALUE_B_MIN>29</VALUE_B_MIN>
8        <VALUE_B_MAX>107</VALUE_B_MAX>
9        <VALUE_C_MIN>154</VALUE_C_MIN>
10       <VALUE_C_MAX>162</VALUE_C_MAX>
11       <CHANNELS>8</CHANNELS>
12     </MODEL>
13   </SYMBOLIC>
14 </COLOR>
15 ...
```

The advantage of this representation is its great flexibility. It enables the usage of various different color models no matter of how many channels it consists. The predicate *Color* is, however, only one useful unit. Another one, supported by the Modality Converter is the predicate *Relation* which is described next.

Relation

In order to be able to model relations in a qualitative Scene Model that exist between objects and locations, a corresponding predicate *Relation* has been implemented. Although these relations are currently not used in the overall architecture of the mobile robot platform as the speech processing units cannot deal with them, yet, relations are internally used for a single object representation. For the use of relations between several objects, the Modality Converter already supports an easy method to transform symbolic names like, e.g., in front of, behind, beneath, and above. This is supported by an XML structure similar to the color representation described in the preceeding paragraph. As the example below shows in line 2, every relation is embedded in a *RELATION* tag. Within such a relation tag, a *SYMBOLIC* tag is given as well, which in this case contains the name “under”. The actual main difference can be seen in lines 4 to 6, where a *HORIZONTAL* tag, a *VERTICAL* tag, and a *DEPTH* tag enclose a mathematical expression, e.g., ’equals’ or ’less’.

```
1  ...
2  RELATION>
3    <SYMBOLIC name="under">
4      <HORIZONTAL>equals</HORIZONTAL>
5      <VERTICAL>less</VERTICAL>
6      <DEPTH></DEPTH>
```

```
7    </SYMBOLIC>
8  </RELATION>
9  ...
```

This described conversion again contains symbolic names, but this notation offers an indisputable advantage. As there are a great deal less mathematical expressions to be considered than there may exist given names for a relation, it is easy to perform a pattern matching in order to select the appropriate computation method. For instance, for the relation 'under', the expression is always 'less' or 'greater' depending on the regarded object. But, the user can have a lot more expressions for that relation even in different languages, e.g., *under, beneath, below, unter, unterhalb, darunter*, and many more. Thus, this kind of transformation tremendously reduces the amount of possible combinations.

The same principles are used for the predicate *Shape* as the following paragraph describes.

Shape

Another important feature to describe an object is the object's shape. Hence, it has been implemented in the Modality Converter as well. The representation of the feature *Shape* is very similar to the one for relations. The only difference is the tag *METHOD* embedded in the *SYMBOLIC* tag, which contains the name of the algorithm used. In this way, only a few algorithms need to be implemented while the user has got the freedom to store various alias names for the object shape, like round, circle, and others. As the actual connection of a symbolic description and an algorithm requires the knowledge of an expert, a couple of connections have already been predefined, mainly based on the suggestions made by Berg in [Ber05], and the Handbook of Pattern Recognition and Computer Vision by Wang and Chen [WC05].

```
1  ...
2  <SHAPE>
3    <SYMBOLIC name="round">
4        <METHOD>Haralick</METHOD>
5    </SYMBOLIC>
6  </SHAPE>
7  ...
```

Although the concrete algorithms are not yet implemented in the Object Attention System, this approach supports a great flexibility for shape recognition. However, as it is not possible to consider all imaginable expressions for different shapes, a method which enables the learning of additional connections needs to be considered as well. For this, the basic idea is as follows. All shapes can principally be assigned to shape primitives, like it is presented in section 4.1 beginning on page 47. Thus, it is imaginable that the robot displays a selection of possible shape primitives on its display, and then, the user selects the most similar visual appearance, e.g., with help of the touch screen interface. As a consequence a new connection has been learned which is even for unexperienced users an easy solution. Such a shape-matching method is, e.g., described by T. Käster in [Käs05] for an approach of intelligent search in an image database.

To sum up, the proposed method to deal with shapes of objects offer an intuitive and convenient solution for a natural Human-Robot Interaction. This is additionally extended in the following by the description of the feature *Size*, which is used by the Object Attention System as well.

Size

The feature type *Size* is used in several distinctive ways within the Object Attention System. Firstly, it is used to determine the distance between the hand and the actual Region-Of-Interest. Secondly, it is used to determine the size of the Region-Of-Interest and, thirdly, it is used to adjust the zoom factor of the object camera (cf. section 4.2 on page 57) in order to get a more detailed view of an object. Unfortunately, the size of an object is usually a highly subjective predicate. While one person calls a ball *normal-sized*, another person can call the ball *small*. So, how can an object size be modeled appropriately? Should it be set in relation to the size of the hand? Or should it be based on size relations to other objects of the same type in the vicinity? Both would make sense, but as these ambiguities demonstrate, the individual determination of the size is influenced by everyone's experiences. Thus, in this thesis the size parameter is determined based on the average size of the user's hand. Nevertheless, as these values are encoded in XML and parsed at runtime, an adaptive adjustment is imaginable. In this way, the user can assign an individual size value for every single learned object. Although this is probably the best solution, it is currently not supported, as the speech processing units cannot deal with anaphoric resolution, i.e., if a sentence clearly names an object and a second, subsequent, sentence refers to this object by, e.g., 'it', indirectly. Hence, a subsequent addition or correction of the size is not possible yet.

The basic structure of the XML fragment below is the same as for all other features as well. The characteristical difference is, thus, only specified in the lines 4 and 5 which contain a minimum and a maximum value for the symbolic size name.

```
1 ...
2 <SIZE>
3   <SYMBOLIC name="large">
4     <VALUE_MIN>20</VALUE_MIN>
5     <VALUE_MAX>40</VALUE_MAX>
6   </SYMBOLIC>
7 </SIZE>
8 ...
```

This value range consequently enables the Object Attention System to use the most appropriate value for a specific computation. For instance, on the one hand, the camera is always zoomed in correspondence to the maximum value in order to ensure that the complete object can be captured. On the other hand, the center of the Region-Of-Interest is based on the mean value of the given value range, as both, the minimum and the maximum value needs to be considered.

A.3 Sound Collector

The basic idea of the Sound Collector developed for the Object Attention System has been presented in chapter 4.5.5 on page 78. Next, a description of the compression algorithms applied within the Sound Collector is given.

In this thesis two different destination formats for sound are considered. The first one is object sound. In this case, the raw PCM-encoded audio file is compressed with the Vorbis [Fou06] audio compression scheme, embedded in an Ogg [Fou06] container file, as these formats are patent and license-free. The Vorbis-encoder was selected as it produces better auditory results than the commonly used proprietary MPEG-1, layer 3 (MP3) format related to two equal input files which are compressed with the same bit rate. The Vorbis compression is well suited for music and sound while it is able to compress the audio file compared to the uncompressed signal to its tenth part usually without subjectively detectable significant information loss in quality.

For applications that concern speech compression, the Sound Collector uses the also patent and license-free Speex codec [JMV06], also embedded in an Ogg container. This codec was selected because *Speex* is specialized on speech compression tasks. It is based on *Code Excited Linear Prediction (CELP)* [JMV06] and, therefore, a Speex-compressed audio file typically provides a 2–4 times higher compression at an equal quality as Vorbis-encoded files. In order to improve the quality of the speech fragment, an automatic gain control and a denoising algorithm is applied in a preprocessing step. Here, the built-in parameters '--agc' and '--denoise' of the 'speexenc' program [JMV06] are used.

A.4 Short-Term Memory

The implemented Short-Term Memory of the Object Attention System is responsible for an accurate maintenance of all interaction-dependent data. Due to its complexity, a typical processing cycle is described to illustrate the internal proceeding within the Short-Term Memory after a query for stored data has been initiated, cf. page 61.

In the data retrieving case, the data is first checked for its validity concerning its age with regard to the computed *Best Before* time. If the data is not valid, the stored entry is deleted immediately, otherwise the data fusion process is initiated. Here, the dialog data is evaluated first, as it can contain a *Deactivate* command if the user aborts an order or the robot has lost the Person-Of-Interest. The reason why this is not already checked during the addition of the data is that even in the current implementation it takes less than 1 second to store, retrieve and evaluate the speech data within the Object Attention System. Thus, a user would have to give two orders, e.g., deactivation and focusing on an object, within one second which is highly unlikely and considering the time for speech processing actually not possible at all. If the Object Attention System receives a deactivation command, it deletes all stored entries in the Short-Term Memory in order to ensure a small memory consumption and a short access time on subsequent interaction data. But, if the received command is different from the deactivation case, the stored input is evaluated in detail which enables the creation of an appropriate object representation for internal use within the Object Attention System. The creation of this structure is discussed in the following, beginning with the speech representation.

Speech Representation

Regarding the speech data representation, at first a distinction is made to determine which dialog model is currently used, the one of I. Toptsis [THH+05] or the

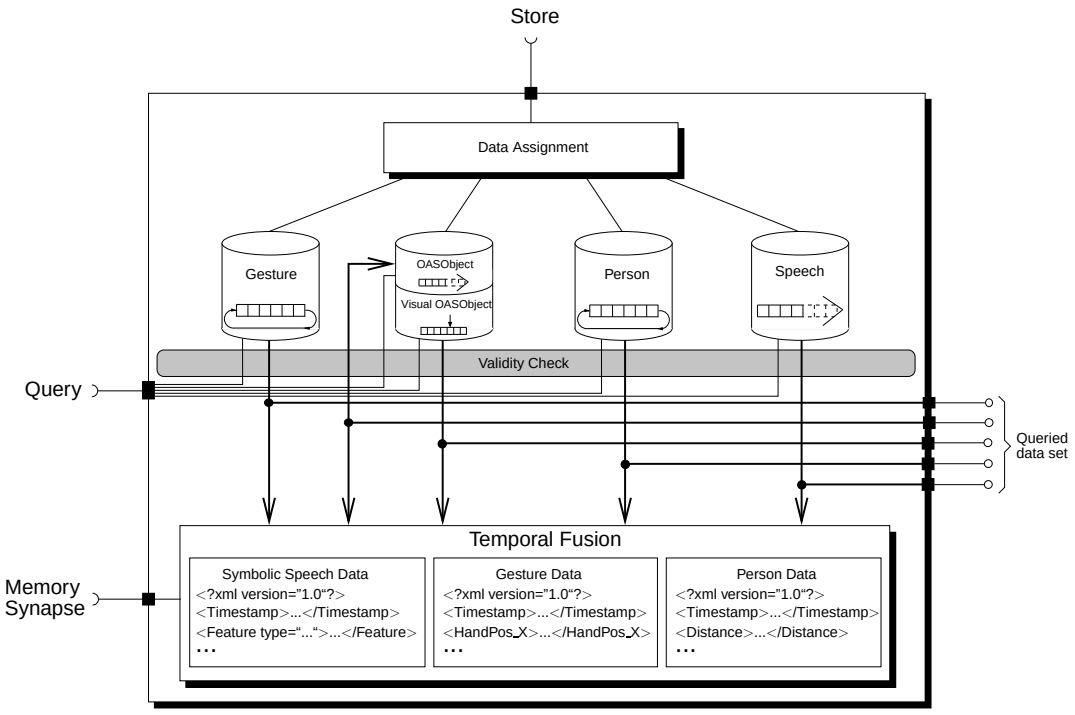


Figure A.1: Schematic illustration of the Short-Term Memory developed for the proposed Object Attention System. Please see the text for a detailed description.

one of S. Li [LHW+05], see also section 3.1 on page 41. Then, all metadata, like timestamp or origin, and object-dependent tags are extracted and depending on their data type (e.g., float, string, ...) transferred into an appropriate structure, a so-called *OASObject*. Such an *OASObject* stores all relevant textual object information as long as it is stored within the Short-Term Memory and, thus, used for internal processing tasks.

Besides the extraction of the data provided by the speech processing units, a confidence value is added for every contained entry as it is described in section 3.1 on page 34. These confidence values (e.g., 85%) are used for later object recognition tasks. As it cannot be assumed that all processed object attributes are all-embracing with regard to all imaginable possible attributes which the user addresses in an interaction, only a selection of attributes useful for object recognition tasks is explicitely stored in a corresponding data field. For all not explicitely processed attributes a flexible vector representation has been chosen in order to be able to store these attribute-value pairs in the long-term object representation. In its current implementation the Object Attention System considers the following attributes: *Object Type*, *Color*, *Consistence* (e.g., liquid, solid, ...), *Haptic* (e.g., soft, hard), *Owner*, *Relation* (e.g., to other objects or locations), *Shape*, *Size*, *Temperature* (symbolic, e.g., hot, as well as numeric), *Article* (definite, indefinite) while the latter one can be used for amounts (e.g., these, this). It is easy to understand that due to the varying amount of specified attributes, the duration of an interaction can vary on a great time scale as well. To sum up, a complex scenario containing a lot of objects while each object contains a number of different

attributes demands for the storage of more utterances than a short conversation about objects. Hence, not only for the attribute list, but also for the overall speech memory another flexible vector representation has been implemented. This will dynamically increase its storage size in dependence of the amount of utterances. These task-dependent settings are therefore regulated by a global configuration that specifies, for instance, the memory duration as well.

After the speech input analysis is complete, the Short-Term Memory analyzes the gesture data.

Gesture Representation

The need for a fast response of the robot in a natural Human-Robot Interaction makes it necessary that the Short-Term Memory stores and processes only temporally appropriate gesture entries. This means that too old and, thus, invalid entries, for those it is unlikely that they belong to an utterance are deleted, cf. section 4.1. Hence, a very fast memory structure has to be used. For the Short-Term Memory this has been realized as ring memory that stores only a limited amount of gesture entries, illustrated in Figure A.1. This enables the memory to deal with very few relevant gesture data sets and, hence, improves the performance of the memory and the interaction, respectively.

After the most appropriate gesture data set has been retrieved from memory, its content is analyzed for the fusion with the speech data. The gesture recognition output contains different data tags depending on the method used. This can either be the result of a skin color-based region tracking or the result of the 3D-Body Model Tracking System, described in section 3.1 on page 36. Hence, based on the *Origin* tags (cf. page 39) it is decided which parsing method is used. Nevertheless, it can happen that the user's gesture has not been recognized. In order to avoid a subsequent failure of the object learning algorithms all attributes, therefore, are initialized with values that cause the robot to lower its camera. In this way, the object camera at least captures the area at a typical height for table tops. Furthermore, the field of view of the object camera is set to a maximum value which enables a larger capturing area.

The gesture is mainly used to determine the Region-Of-Interest which is assumed to contain the referenced object. As an outcome of the gesture processing, a set of data, e.g., including 3D-coordinates for the user's hand and the referenced location, becomes available that is added to a special *body coordinate* object. Such an additional structure is useful as, for instance, the user's head position is usually not relevant for the location of an object. Furthermore, this ensures to keep the actual *OASObject* small in memory consumption, as it does only need to store the real object location which is computed later during the processing cycle.

Now that the most appropriate gesture data has been parsed and assigned to the currently processed utterance, the information about the Person-Of-Interest is analyzed next in the Short-Term Memory.

Person Representation

The information about the current interacting user is provided by the Person Attention and Tracking System, cf. section 3.1 on page 34. Due to its relatively high

frequency of up to 5 Hertz, the data is stored in a ring buffer as well. That way, only the most recent positions of the user are considered which results in a fast access time and a low memory consumption.

During the parsing and temporal assignment of person data to the processed utterance, a comparison between data of the face identification and the given speech information is performed. In case the user did not state his name, the corresponding data field is set up with the name from the face recognition. But, this is only assigned if the face recognition was successful, otherwise the value *unknown* is assigned to the data field *owner*, see page 80. Furthermore, the distance from the user is extracted in order to be able to compute the correction factor, described in equation 3.1 on page 38. Summarizing, the extracted person data is assigned to two different structures. Firstly, the owner information to an OASObject and, secondly, the person distance to a body coordinate object.

After all modalities have been processed and assigned to an OASObject or a body coordinate object, the next section gives a detailed description of the object representation itself.

Object Representation

According to the requirements, objects are represented in two different forms and, thus, stored in two different memory structures. The first one is the already mentioned OASObject. The corresponding memory infrastructure for those objects is implemented as a resizable vector as its content varies in dependency of already learned objects that are used for an object recognition task.

The second memory type for objects is a specialization of an OASObject, a so-called *Visual OASObject*. Its purpose is to manage visual object data, like, e.g., the center of mass in image coordinates. For this representation, a map container has been chosen as it allows a direct access to each element of the container.

B. Evaluation Tables

This chapter contains a detailed representation of the results that have been achieved during the evaluation of the Object Attention System. They have not been directly included in chapter 6 as they would minimize the clarity of the statements for the conducted experiments.

Evaluation table for the Determination of the Region-Of-Interest

A detailed overview about all 576 calculated Region-Of-Interest values is presented in table B.1 on page 131. They are related to the summarized values that have been presented in chapter 6.1 on page 105. In detail, the table shows the 486 calculated position values for the Regions-Of-Interest while the 90 calculated values for the minimum, the mean, and the maximum error values for each object is presented separately.

In the upper part of the table, the five objects used, in particular, a red bottle, a green toy crocodile, a blue cup, a yellow lemon, and a pink ball (cf. Figure 6.1 on page 102) are denoted on the abscissae. The ordinate is separated by the four participants. The first and the third person represent the unexperienced users and the second and fourth user are the developers of the 3D-Body Tracking System and the Gesture Recognition. Each row denoting the participants is divided into the four sequences that are described in the evaluation chapter on page 103. As the table illustrates, not for all runs exist 3 entries for an estimated object location. In these cases the gesture has not been successfully recognized and, thus, not sent to the Object Attention System. As a consequence, these gestures could not be evaluated.

For a direct comparison with the manually measured positions (cf. green values at bottom of the table), the values that either match or if no exact match exists, the most similar values are denoted in green color. The red-colored values are related to the maximum error that has been calculated relative to the manually measured object location.

Experimental runs per person	Person	Run	Objects and their estimated positions														
			Green crocodile			Blue cup			Pink ball			Yellow lemon			Red bottle		
			Height [m]	Distance [m]	∡ [°]	Height [m]	Distance [m]	∡ [°]	Height [m]	Distance [m]	∡ [°]	Height [m]	Distance [m]	∡ [°]	Height [m]	Distance [m]	∡ [°]
	1	1	0.78	2.39	-15.6	0.64	2.18	-8.6	0.77	2.24	3.2	0.77	2.39	6.3	0.75	2.47	7.0
			0.83	2.42	-14.1	0.77	2.23	-7.6	0.73	2.19	3.4	0.82	2.40	5.7	0.76	2.22	2.4
			0.80	2.42	-17.0										0.68	2.43	7.6
		2	0.78	2.37	-14.5	0.71	2.21	-7.0	0.76	2.26	2.4	0.70	2.24	3.3	0.69	2.45	8.3
			0.74	2.33	-15.8	0.70	2.25	-10.7	0.75	2.28	8.3				0.85	2.29	5.3
		3	0.85	2.40	-15.6	0.71	2.24	-8.4	0.71	2.19	2.4	0.67	2.27	5.0	0.79	2.55	10.6
			0.74	2.44	-15.9	0.63	2.20	-8.9	0.76	2.22	5.2	0.71	2.27	2.8			
		4	0.76	2.38	-14.6	0.72	2.22	-6.7	0.65	2.21	6.0	0.72	2.34	4.4	0.63	2.47	9.0
			0.73	2.32	-15.4	0.65	2.18	-9.3	0.69	2.17	6.2	0.70	2.23	1.8	0.74	2.44	7.4
	2	1	0.76	2.28	-14.9	0.69	2.08	-7.3	0.69	2.03	1.4	0.68	2.14	4.1	0.72	2.42	7.2
						0.72	2.13	-7.7							0.63	2.36	7.8
						0.69	2.10	-11.8									
		2	0.78	2.26	-13.1	0.77	2.16	-7.2	0.74	2.09	3.3	0.72	2.15	3.8	0.68	2.34	7.6
			0.77	2.27	-14.5	0.67	2.06	-8.5	0.78	2.11	4.1	0.75	2.19	3.9	0.72	2.28	6.2
		3	0.81	2.29	-13.5	0.72	2.10	-7.0	0.72	2.07	0.8	0.65	2.10	4.4	0.66	2.36	11.5
			0.72	2.18	-15.5	0.73	2.20	-9.2	0.76	2.12	1.8	0.71	2.21	6.4	0.76	2.31	8.6
						0.67	2.12	-13.4							0.66	2.33	9.6
		4	0.77	2.25	-13.5	0.70	2.09	-6.7	0.74	2.11	1.4	0.69	2.20	6.5	0.64	2.38	11.9
			0.80	2.32	-14.8	0.83	2.27	-9.6	0.80	2.20	2.8	0.76	2.25	6.9	0.74	2.36	12.2
															0.68	2.38	10.5
	3	1	0.62	2.06	-13.6	0.55	2.00	-9.0	0.64	2.04	4.4	0.51	1.99	4.7	0.51	2.24	9.1
						0.53	1.90	-7.5	0.59	1.98	5.6	0.56	2.03	4.8	0.59	2.30	9.4
															0.49	2.21	7.7
		2	0.63	2.14	-14.5	0.53	1.94	-9.5	0.62	2.02	6.1	0.56	2.05	5.2	0.58	2.33	9.9
			0.64	2.14	-14.1	0.55	2.02	-8.2	0.59	2.03	3.7	0.60	2.06	5.2	0.61	2.38	10.1
															0.51	2.30	10.2
		3	0.63	2.15	-15.2	0.48	1.95	-7.9	0.59	1.98	4.8	0.58	2.05	4.0	0.60	2.37	8.6
			0.55	2.11	-15.1				0.48	1.97	4.6				0.59	2.17	7.1
		4	0.66	2.19	-15.4	0.54	1.99	-10.7	0.58	1.98	5.4	0.53	1.98	2.1	0.59	2.22	7.6
			0.57	2.12	-14.6	0.51	1.98	-9.4	0.55	2.01	4.5	0.45	1.97	2.7	0.66	2.07	4.9
															0.41	2.19	6.4
	4	1	0.73	2.21	-14.5	0.57	2.03	-8.4	0.59	2.01	3.9	0.63	2.22	5.2	0.63	2.41	8.6
			0.76	2.16	-12.4	0.54	1.97	-10.2				0.58	2.20	5.2	0.59	2.16	8.7
			0.72	2.29	-19.1										0.56	2.33	7.7
		2	0.68	2.24	-16.0	0.54	2.03	-8.1	0.60	2.02	4.2	0.62	2.30	6.5	0.59	2.42	9.3
			0.65	2.26	-16.9	0.63	2.16	-11.7	0.63	2.16	7.0	0.52	2.19	4.6	0.65	2.39	9.2
															0.56	2.43	10.6
		3	0.71	2.21	-16.5	0.61	2.10	-9.0	0.68	2.12	1.6	0.56	2.25	6.9	0.55	2.37	8.8
			0.64	2.24	-18.2	0.53	2.01	-9.8	0.65	2.06	2.4	0.52	2.17	6.9	0.61	2.21	9.8
		4	0.64	2.13	-15.1	0.54	1.97	-8.1	0.48	2.06	1.4	0.56	2.27	6.9	0.61	2.45	10.8
			0.65	2.15	-15.3				0.54	2.10	2.0	0.51	2.18	5.5	0.49	2.20	8.8
Reference measurement			0.75	2.17	-13.6	0.75	1.95	-7.5	0.75	2.01	0.0	0.75	1.93	7.1	0.75	2.15	19.7
Min. error [%]			1.3	0.5	0.0	2.7	0.0	0.0	0.0	0.0	0.8	0.0	2.1	2.8	0.0	0.5	38.1
Min. error [m/°]			0.01	0.01	0.0	0.02	0.0	0.0	0.0	0.0	0.8	0.0	0.04	0.2	0.0	0.01	7.5
Mean error [%]			4.0	0.4	11.8	14.7	7.7	17.3	12.0	4.5	8.3	16.0	13.0	31.0	16.0	8.4	56.3
Mean error [m/°]			-0.03	0.08	1.6	-0.11	0.15	1.3	-0.09	0.09	3.8	-0.12	0.25	-2.2	-0.12	0.18	-11.1
Max. error [%]			26.7	12.4	40.4	36.0	16.4	78.7	36.0	13.4	8.3	40.0	24.4	74.6	54.7	18.6	87.8
Max. error [m/°]			-0.20	0.27	5.5	-0.27	0.32	5.9	-0.27	0.27	8.3	-0.30	+0.47	-5.3	-0.34	0.40	-17.3

Table B.1: Results of experiments for the calculation of Regions-Of-Interest. Please see text on page 129 for a detailed description.

This page intentionally left blank.

Evaluation table for the qualitative measurement of object-related depth values

The details on the calculated depth-based object positions are presented in this section, cf. section 6.3 on page 110. The achieved results are, therefore, given in table B.2 on the next page. In the most left column, the different evaluation objects are listed. Additionally, the column contains entries for the calculated mean value and the resulting standard deviation in relation to each of the three runs performed. The column titled $||\vec{o}_l||$ and $||\vec{o}_r||$ show the calculated distance values between the robot camera and the object for the left and the right object, respectively. This value has manually been measured and amounts 1.00 m. The last column $||\vec{o}_r - \vec{o}_l||$ denotes the relative distances between the two objects. The manually measured value for this relation amounts 30 cm.

Object l	Object r	$\|\|\vec{o}_l\|\|$	$\|\|\vec{o}_r\|\|$	$\|\|\vec{o}_r - \vec{o}_l\|\|$
White cup	Black cup	1.00	1.26	0.35
Black cup	White cup	1.02	1.16	0.27
Lemon	Bottle	1.00	1.19	0.29
Bottle	Lemon	1.03	1.18	0.27
Remote	Clock	0.94	1.48	0.59
Clock	Remote	0.94	1.27	0.41
Mould	Candle	1.08	1.32	0.34
Candle	Mould	1.05	1.17	0.26
Apple	Cherry	1.11	1.29	0.29
Cherry	Apple	1.31	1.33	0.26
	Mean value	1.048	1.265	0.333
	Standard deviation	0.1067	0.0983	0.1025
White cup	Black cup	0.94	1.17	0.32
Black cup	White cup	0.95	1.09	0.25
Lemon	Bottle	0.99	1.17	0.29
Bottle	Lemon	1.04	1.28	0.34
Remote	Clock	0.95	1.16	0.30
Clock	Remote	0.94	1.20	0.34
Mould	Candle	1.02	1.17	0.28
Candle	Mould	1.00	1.18	0.28
Apple	Cherry	1.00	1.25	0.35
Cherry	Apple	0.98	1.23	0.34
	Mean value	0.981	1.190	0.309
	Standard deviation	0.0351	0.0533	0.0338
White cup	Black cup	0.83	0.99	0.25
Black cup	White cup	0.82	0.94	0.22
Lemon	Bottle	0.82	0.99	0.25
Bottle	Lemon	0.82	1.01	0.27
Remote	Clock	0.85	0.99	0.24
Clock	Remote	0.84	0.93	0.20
Mould	Candle	0.85	1.00	0.24
Candle	Mould	0.84	1.02	0.26
Apple	Cherry	0.85	1.03	0.26
Cherry	Apple	0.87	1.03	0.24
	Mean value	0.839	0.993	0.243
	Standard deviation	0.0166	0.0343	0.0206

Table B.2: Evaluation results for the determination of object positions. All values are specified in [m]. The table has been adapted from [Köl06].

Bibliography

[AC05] K. O. Arras and D. Cerqui. Do we want to share our lives and bodies with robots? A 2000-people survey. Technical report, Autonomous Systems Lab. Swiss Federal Institute of Technology, EPFL, Lausanne, Switzerland, 2005.

[ALTB05] M. Berlin A. L. Thomaz and C. Breazeal. An Embodied Computational Model of Social Referencing. In *Proc. of 14th IEEE Workshop on Robot and Human Interactive Communication (RO-MAN)*, Nashville, TN, 2005.

[Apa06a] Apache Software Foundation. Log4cxx, 2006. http://logging.apache.org/log4cxx/ (Date: May 31, 2007).

[Apa06b] Apache Software Foundation. Log4j, 2006. http://logging.apache.org/log4j/docs/index.html (Date: May 31, 2007).

[Bac04] G. Backer. *Modellierung visueller Aufmerksamkeit im Computer-Sehen: Ein zweistufiges Selektionsmodell für ein Aktives Sehsystem*. Dissertation, University of Hamburg, Germany, 2004.

[BBG+04] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda. Humanoid Robots as Cooperative Partners for People. *Int. Journal of Humanoid Robots*, 2004.

[BBM05] A. C. Berg, T. L. Berg, and J. Malik. Shape Matching and Object Recognition using Low Distortion Correspondence. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005.

[BBS04] C. Bauckhage, E. Braun, and G. Sagerer. From Image Features to Symbols and Vice Versa – Using Graphs to Loop Data- and Model-Driven Processing in Visual Assembly Recognition. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 18(3):497–517, 2004.

[BC91] S. Baron-Cohen. *Precursors to a theory of mind: Understanding attention in others*, pages 233–251. Blackwell Press, Oxford, UK, 1991.

[BC95] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.

[BE04] M. Björkman and J.O. Eklundh. Attending, Foveating and Recognizing Objects in Real World Scene. In *British Machine Vision Conf. (BMVC)*, 2004.

[Ber05] A. C. Berg. *Shape Matching and Object Recognition*. PhD thesis, Computer Science Division, U.C. Berkeley, 2005.

[BG99] R. Bischoff and V. Graefe. Integrating Vision, Touch and Natural Language in the Control of a Situation-Oriented Behavior-Based Humanoid Robot. In *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, Tokyo, Japan, 1999.

[BG02] R. Bischoff and V. Graefe. Demonstrating the Humanoid Robot *HERMES* at an Exhibition: A Long-Term Dependability Test. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems; Workshop on Robots at Exhibitions*, Lausanne, Switzerland, 2002.

[BGW+06] B. Bakker, C. Le Goater, M. Welz, L. Owen, S. Ostlind, M. Harkema, U. Jäger, W. Stroebel, G. Scott, T. Cheung, B. B. Boerner A. Tapaccos, P. Pizarro, D. Resnick, A. Ingram, A. Anderson, and E. Martin. Log for C++, 2006. http://log4cpp.sourceforge.net/ (Date: May 31, 2007).

[BJ91] G. E. Butterworth and N. L. M. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[BMS+05] C. Burghart, R. Mikut, R. Stiefelhagen, T. Asfour, H. Holzapfel, P. Steinhaus, and R. Dillmann. A Cognitive Architecture for a Humanoid Robot: A First Approach. In *Proc. of 5th IEEE/RAS Int. Conf. on Humanoid Robots*, Tsukuba, Japan, 2005.

[Bra06] E. Braun. *A Framework for Integrating Object Recognition Strategies*. Dissertation, Bielefeld University, Faculty of Technology, 2006.

[BRB+04] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *Proc. European Conf. on Computer Vision (ECCV)*, 2004.

[BS99] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1146–1151, Stockholm, Sweden, 1999.

[BSD03] R. Becher, P. Steinhaus, and R. Dillmann. Interactive Object Modelling for a Humanoid Service Robot. In *Proc. of 3rd IEEE Int. Conf. on Humanoid Robots*, 2003.

[BSZD06] R. Becher, P. Steinhaus, R. Zöllner, and R. Dillmann. Design and Implementation of an Interactive Object Modelling System. In *Proc. Robotik/IRS*, 2006.

[BT98] S. Birchfield and C. Tomasi. Depth Discontinuities by Pixel-to-Pixel Stereo. In *Proc. of the Int. Conf. on Computer Vision (ICCV)*, pages 1073–1080, Bombay, India, 1998.

[CD02] B. Caputo and G. Dorko. How to combine color and shape information for 3d object recognition: kernels do the trick, 2002.

[COG06] COGNIRON. The Cognitive Robot Companion, 2006. (FP6-IST-002020), http://www.cogniron.org (Date: May 31, 2007).

[CS03] S. Coradeschi and A. Saffiotti. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2–3):85–96, 2003.

[Dic99] S. Dickinson. Object Representation and Recognition. In E. Lepore and Z. Pylyshyn, editors, *Rudgers University Lectures on Cognitive Science*, pages 172–207. Basil Blackwell publishers, 1999.

[DWK+05] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. L. Koay, and I. Werry. What is a Robot Companion – Friend, Assistant or Butler? In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, 2005.

[Eme00] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604, 2000.

[FB81] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[fE04] World Health Organization. Regional Office for Europe. Age pyramid for Germany, 2004. http://www.euro.who.int/eprise/main/WHO/Progs/CHHDEU/annex/20041126_1 (Date: May 31, 2007).

[Fin99] G. A. Fink. Developing HMM-based Recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

[FKH+05] J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, and G. Sagerer. A Flexible Infrastructure for the Development of a Robot Companion with Extensible HRI-Capabilities. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3419–3425, Barcelona, Spain, 2005.

[FKL+03] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. 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):133–147, 2003.

[FLT06] FLTK. Fast Light Toolkit (FLTK), 2006. http://www.fltk.org (Date: May 31, 2007).

[FND03] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

[Fou06] Xiph.org Foundation. Ogg Vorbis compression format, 2006. http://www.vorbis.com/ (Date: May 31, 2007).

[fPuA05] Fraunhofer-Institut fur Produktionstechnik und Automatisierung. Desire – Deutsche Servicerobotik Initiative, 2005. http://www.ais.fraunhofer.de/BE/DESIRE/index.html (Date: May 31, 2007).

[Fri03] J. Fritsch. *Vision-based Recognition of Gestures with Context*. Dissertation, Bielefeld University, Faculty of Technology, 2003.

[FS95] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *European Conf. on Computational Learning Theory*, pages 23–37, 1995.

[FWS05] J. Fritsch, B. Wrede, and G. Sagerer. Bringing it all together: Integration to study embodied interaction with a robot companion. In *AISB Symposium – Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*, Hatfield, England, 2005.

[GHS04] B. Graf, M. Hans, and R. D. Schraft. Mobile Robot Assistants – Issues for Dependable Operation in Direct Cooperation With Humans. *IEEE Robotics & Automation Magazine*, 11(2):67–77, 2004.

[GNS+02] S. S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori. Multi-Modal Interaction of Human and Home Robot in the Context of Room Map Generation. *Autonomous Robots*, 13(2):169–184, 2002.

[GRM+06] H.-M. Gross, J. Richarz, S. Mueller, A. Scheidig, and C. Martin. Probabilistic Multi-modal People Tracker and Monocular Pointing Pose Estimator for Visual Instruction of Mobile Robot Assistants. In *Proc. IEEE World Congress on Computational Intelligence & Int. Joint Conf. on Neural Networks (IJCNN)*, pages 8325–8333, 2006.

[GVH03] B. P. Gerkey, R. T. Vaughan, and A. Howard. The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems. In *Proc. Int. Conf. on Advanced Robotics*, pages 317–323, Coimbra, Portugal, 2003.

[GWOG99] L. Gonçalves, D. Wheeler, A. Oliveira, and R. Grupen. Towards a Framework for Robot Cognition. In *Proc. of the IEEE Int. Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, Monterey, CA, USA, 1999.

[GYJX99] D. Guo, X. M. Yin, Y. Jin, and M. Xie. Efficient Gesture Interpretation for Gesture-based Human-Service Robot Interaction (FSR). In *Proc. of the Int. Conf. on Field and Service Robotics*, Pittsburgh, Pennsylvania, 1999.

[Har94a] R. I. Hartley. Self-Calibration from Multiple Views with a Rotating Camera. In *Proc. of 3rd European Conf. of Computer Vision (ECCV)*, pages 471–478, Stockholm, Sweden, 1994.

[Har94b] R. I. Hartley. Self-Calibration from Two Views, 1994. http://users.rsise.anu.edu.au/~hartley/ (Date: May 31, 2007).

[HFS04] N. Hofemann, J. Fritsch, and G. Sagerer. Recognition of Deictic Gestures with Context. In C. E. Rasmussen, H. H. Bülthoff, M. A. Giese, and B. Schölkopf, editors, *DAGM*, volume 3175 of *Lecture Notes in Computer Science*, pages 334–341, Heidelberg, Germany, 2004. Springer.

[HGS02] M. Hans, B. Graf, and R. D. Schraft. Robotic home assistant care-O-bot: Past-Present-Future. In *Proc. IEEE Int. Workshop Robot and Human Interactive Communication (RO-MAN)*, pages 407–411, Bordeaux, Paris, France, 2002.

[HHFS05] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A Multi-Modal Object Attention System for a Mobile Robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1499–1504, Edmonton, Alberta, Canada, 2005.

[HHH+04] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON – The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32, Stuttgart, Germany, 2004. Fraunhofer IRB.

[HNS04] H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D Pointing Gestures. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, pages 175–182, State College, PA, USA, 2004.

[Hof06] N. Hofemann. *Videobasierte Handlungserkennung für die natürliche Mensch-Maschine-Interaktion*. Dissertation, Bielefeld University, Faculty of Technology, 2006.

[Hoh05] S. Hohenner. *Automatische Spracherkennung für agierende Systeme*. Dissertation, Bielefeld University, Faculty of Technology, 2005.

[HSF+05] M. Hackel, S. Schwope, J. Fritsch, B. Wrede, and G. Sagerer. A Humanoid Robot Platform Suitable for Studying Embodied Interaction. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 56–61, Edmonton, Alberta, Canada, 2005.

[HW06a] N. Hawes and J. Wyatt. Towards Context-Sensitive Visual Attention. In *Proc. of the 2nd Int. Cognitive Vision Workshop (ICVW)*, Graz, Austria, 2006.

[HW06b] S. Hüwel and B. Wrede. Situated Speech Understanding for Robust Multi-Modal Human-Robot Communication. In *Proc. of the Int. Conf. on Computational Linguistics (COLING/ACL)*. ACL Press, 2006.

[HWBR06] J. Hois, M. Wünstel, J. A. Bateman, and T. Röfer. Dialog-Based 3D-Image Recognition Using a Domain Ontology. In *Spatial Cognition V*, Lecture Notes in Artificial Intelligence. Springer, 2006.

[HZ04] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edt. edition, 2004.

[IK00] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12):1489–1506, 2000.

[IKN98] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998.

[int06] intel®. OpenCV (Open Source Computer Vision) library, 2006. http://www.intel.com/technology/computing/opencv/, http://opencvlibrary.sourceforge.net/ (Date: May 31, 2007).

[IOI+01] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu. Robovie: An interactive humanoid robot. *Int. Journal Industrial Robotics*, 28(6):498–503, 2001.

[JMV06] Xiph.org Foundation J.-M. Valin. Speex: A Free Codec For Free Speech, 2006. http://www.speex.org/ (Date: May 31, 2007).

[Kah96] R. E. Kahn. *Perseus: An extensible Vision System for Human-Machine Interaction*. PhD thesis, The University of Chicago, Chicago, Illinois, USA, 1996.

[Käs05] T. Käster. *Intelligente Bildersuche durch den Einsatz inhaltsbasierter Techniken*. Dissertation, Bielefeld University, Faculty of Technology, 2005.

[KBCE05] D. Kragic, M. Björkman, H. I. Christensen, and J.-O. Eklundh. Vision for Robotic Object Manipulation in Domestic Settings. *Robotics and Autonomous Systems*, 52:85–100, 2005.

[KD05] Y. Keselman and S. Dickinson. Generic Model Abstraction from Examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1141–1156, 2005.

[Ken04] A. Kendon. *Gesture*. Cambridge Univ. Press, Cambridge [u.a.], 2004.

[KFS04] M. Kleinehagenbrock, J. Fritsch, and G. Sagerer. Supporting Advanced Interaction Capabilities on a Mobile Robot with a Flexible Control System. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, volume 3, pages 3649–3655, Sendai, Japan, 2004.

[KG01] L. Kopp and P. Gardenfors. Attention as a Minimal Criterion of Intentionality in Robotics. Technical Report 89, Lund University of Cognitive Studies, Lund, Sweden, 2001.

[KH04] F. Kaplan and V. Hafner. The Challenges of Joint Attention. In *Proc. 4th Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 67–74, Genoa, Italy, 2004.

[Kha98] Z. Khan. Attitudes towards intelligent service robots. Technical report, NADA, 1998.

[KK05] J. D. Kelleher and G.-J. M. Kruijff. A context-dependent model of proximity in physically situated environments. In *Proc. of the ACL-SIGSEM workshop The Linguistic Dimension of Prepositions*, Colchester, England, 2005.

[KKBL06] G. Kruijff, J. Kelleher, G. Berginc, and A. Leonardis. Structural descriptions in human-assisted robot visual learning. In *Proc. of 1st ACM SIGCHI/SIGART Conf. on Human Robot Interaction*, pages 343–344, Salt Lake City, Utah, 2006.

[KKH06] G.-J. M. Kruijff, J. D. Kelleher, and N. Hawes. Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue. In E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, editors, *Perception and Interactive Technologies: Int. Tutorial and Research Workshop, PIT*, volume 4021, pages 117–128, Kloster Irsee, Germany, 2006. Springer.

[Kle05] M. Kleinehagenbrock. *Interaktive Verhaltenssteuerung für Robot Companions*. Dissertation, Bielefeld University, Faculty of Technology, 2005.

[KLP+06] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*. Mouton de Gruiter, 2006.

[KLT04] H. Kim, B. Lau, and J. Triesch. Adaptive Object Tracking with an Anthropomorphic Robot Head. In *Proc. of 8th Int. Conf. on the Simulation of Adaptive Behavior (SAB)*, Los Angeles, USA, 2004.

[Köl06] M. Köllmann. Positionsakquisition mit einem anthropomorphen Roboterkopf. Master's thesis, Bielefeld University, Faculty of Technology, 2006.

[Kop03] S. Kopp. *Synthese und Koordination von Sprache und Gestik für virtuelle multimodale Agenten*. PhD thesis, Bielefeld University, Faculty of Technology, 2003.

[KS04] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *IEEE Proc. of Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 506–513, 2004.

[KWK05] S. Kirstein, H. Wersing, and E. Körner. Rapid Online Learning of Objects in a Biologically Motivated Recognition Architecture. In *27th Pattern Recognition Symposium DAGM*, pages 301–308. Springer, 2005.

[KY01] H. Kozima and H. Yano. A Robot that Learns to Communicate with Human Caregivers. In *Proc. Intl. Workshop on Epigenetic Robotics*, Lund, Sweden, 2001.

[Lan05] S. Lang. *Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter*. Dissertation, Bielefeld University, Faculty of Technology, 2005.

[Les94] A. M. Leslie. ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. Hirschfeld and S. Gelman, editors, *Mapping the mind: Domain specificity in cognition and culture*, pages 119–148, New York, 1994. Cambridge University Press.

[Lew95] J. P. Lewis. Fast Template Matching. In *Proc. Conf. on Vision Interface*, pages 120–123, 1995.

[LHW+05] S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, pages 151–158, Trento, Italy, 2005. ACM Press.

[LLS04] B. Leibe, A. Leonardis, and B. Schiele. Combined Object Categorization and Segmentation with an Implicit Shape Model. In *ECCV Workshop on Statistical Learning in Computer Vision*, Prague, 2004.

[Löm04] F. Lömker. *Lernen von Objektbenennungen mit visuellen Prozessen*. Dissertation, Bielefeld University, Faculty of Technology, 2004.

[Löm06] F. Lömker. iceWing, an Integrated Communication Environment Which Is Not Gesten, 2006. http://icewing.sf.net (Date: May 31, 2007).

[Low04] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 20(2):91–110, 2004.

[LPD+01] K. Lay, E. Prassler, R. Dillmann, G. Grunwald, M. Hägele, G. Lawitzky, A. Stopp, and W. von Seelen. MORPHA: Communication and Interaction with Intelligent, Anthropomorphic Robot Assistants. In *Tagungsband Statustage Leitprojekte Mensch-Technik-Interaktion in der Wissensgesellschaft*, Saarbrücken, Germany, 2001.

[LWHF06] F. Lömker, S. Wrede, M. Hanheide, and J. Fritsch. Building Modular Vision Systems with a Graphical Plugin Environment. In *Proc. of IEEE Int. Conf. on Vision Systems*, St. Johns University, Manhattan, New York City, USA, 2006.

[Mat01] M. J. Matarić. Learning in Behavior-Based Multi-Robot Systems: Policies, Models, and Other Agents. *Cognitive Systems Research, Special issue on Multi-disciplinary studies of multi-agent learning*, 2(1):81–93, 2001.

[May79] P. S. Maybeck. *Stochastic models, estimation, and control*. Academic Press, 1979.

[McN92] D. McNeill. *Hand and Mind*. University of Chicago Press, Chicago, Illinois, 1992.

[MFR+02] P. McGuire, J. Fritsch, H. Ritter, J. J. Steil, F. Röthling, G. A. Fink, S. Wachsmuth, and G. Sagerer. Multi-Modal Human-Machine Communication for Instructing Robot Grasping Tasks. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1082–1089, 2002.

[MI05] J. Mukai and M. Imai. Maintenance and Drift of Attention in Human-Robot Communication. In *Int. Conf. on Informatics in Control, Automation and Robotics*, 2005.

[Möl05] B. Möller. *Ein Ansatz zur ikonischen Repräsentation von Bilddaten aktiver Kameras*. PhD thesis, Institute of Computer Science, Martin-Luther University Halle-Wittenberg, Halle (Saale), Germany, 2005.

[MOR99] MORPHA. MORPHA – Intelligent Robot Assistants, 1999. http://www.morpha.de/ (Date: May 31, 2007).

[MPH+05] B. Möller, S. Posch, A. Haasch, J. Fritsch, and G. Sagerer. Interactive Object Learning for Robot Companions using Mosaic Images. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 371–376, Edmonton, Alberta, Canada, 2005.

[Nag04] Y. Nagai. *Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics*. PhD thesis, Osaka University, Osaka, Japan, 2004.

[Nag05a] Y. Nagai. Learning to Comprehend Deictic Gestures in Robots and Human Infants. In *Proc. of the 14th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*, pages 217–222, 2005.

[Nag05b] Y. Nagai. The Role of Motion Information in Learning Human-Robot Joint Attention. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 2081–2086, Barcelona, Spain, 2005.

[NHMA03] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[NM01] M. Nicolescu and M. J. Mataric. Learning and Interacting in Human-Robot Domains. *Special Issue of IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 31(5):419–430, 2001.

[OI00] T. Ono and M. Imai. Reading a Robot's Mind: A Model of Utterance Understanding Based on the Theory of Mind Mechanism. In *AAAI/IAAI*, pages 142–148, 2000.

[Ora06] Oracle. Berkely DB XML, 2006. http://www.oracle.com/database/berkeley-db/index.html (Date: May 31, 2007).

[REZ+02] O. Rogalla, M. Ehrenmann, R. D. Zoellner, R. Becher, and R. Dillmann. Using Gesture and Speech Control for Commanding a Robot Assistant. In *Proc. of the 11th IEEE Int. Workshop on Robot and Human interactive Communication (RO-MAN)*, Berlin, Germany, 2002.

[RKB04] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. In *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[RMSG06] J. Richarz, C. Martin, A. Scheidig, and H.-M. Gross. There you go! - Estimating Pointing Gestures in Monocular Images for Mobile Robot Control. In *Proc. of IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 546–551, Hatfield, UK, 2006.

[RP97] D. Roy and A. Pentland. Multimodal Adaptive Interfaces. Technical report, MIT Media Lab, USA, 1997.

[Sae05] M. Saerbeck. Lernen unbekannter Objekte durch Analyse visueller Merkmale. Master's thesis, Bielefeld University, Faculty of Technology, 2005.

[Sca02] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24, 2002.

[Sca03] B. Scassellati. Investigating models of social development using a humanoid robot. In *Int. Joint Conf. on Neural Networks (IJCNN)*, Portland, OR, 2003.

[Sch01] B. J. Scholl. Objects and Attention. *Cognition*, 80(1–2):1–46, 2001.

[SF03] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[SGW+06] J. J. Steil, M. Goetting, H. Wersing, E. Körner, and H. Ritter. Adaptive scene-dependent filters for segmentation and online learning of visual objects. *Neurocomputing*, 2006.

[SHFS06] T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer. Human-like Person Tracking with an Anthropomorphic Robot. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1286–1292, Orlando, Florida, 2006.

[SIKM02] M. Strobel, J. Illmann, B. Kluge, and F. Marrone. Gesture recognition in a spatial context for commanding a domestic service robot, 2002.

[SKD+06] A. Shokoufandeh, Y. Keselman, F. Demirci, D. Macrini, and S. Dickinson. *Many-to-Many Feature Matching in Object Recognition*, chapter Cognitive Vision Systems: Sampling the Spectrum of Approaches, pages 107–125. Springer, Berlin, Germany, 2006.

[SKF06] J. Schmidt, B. Kwolek, and J. Fritsch. Kernel Particle Filter for Real-Time 3D Body Tracking in Monocular Color Images. In *Proc. of IEEE Automatic Face and Gesture Recognition*, pages 567–572, Southampton, UK, 2006.

[SKI+05] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. Three-Layered Draw-Attention Model for Humanoid Robots with Gestures and Verbal Cues. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 2140–2145, Edmonton, Alberta, Canada, 2005.

[Sow05] T. Sowa. *Understanding Coverbal Iconic Gestures in Shape Description*. PhD thesis, Bielefeld University, Faculty of Technology, 2005.

[Spe05] T. Spexard. Aufmerksamkeitsgesteuerte Interaktion mit einem anthropomorphen Roboter. Master's thesis, Bielefeld University, Faculty of Technology, 2005.

[SRHR04] J. Steil, F. Röthling, R. Haschke, and H. Ritter. Situated robot learning for multi-modal instruction and imitation of grasping. *Robotics and Autonomous Systems*, Special Issue on "Robot Learning by Demonstration"(47):129–141, 2004.

[SW06] J. J. Steil and H. Wersing. Recent Trends in Online Learning for Cognitive Robotics. In *Proc. ESANN*, pages 77–87, 2006.

[SWSK05] G. Schneider, H. Wersing, B. Sendhoff, and E. Körner. Evolutionary optimization of a hierarchical object recognition model. *IEEE Trans. Systems, Man, Cybernetics, Part B: Cybernetics*, 35(3):426–437, 2005.

[TE05] J. Triesch and C. Eckes. *Handbook of Pattern Recognition and Computer Vision. Object Recognition with Deformable Feature Graphs: Faces, Hands, and Cluttered Scenes*, chapter 4.5, pages 461–480. World Scientific Publishing Co., 3rd edition, 2005.

[THH+05] I. Toptsis, A. Haasch, S. Hüwel, J. Fritsch, and G. A. Fink. Modality Integration and Dialog Management for a Robotic Assistant. In *Proc. European Conf. on Speech Communication and Technology*, Lisboa, Portugal, 2005.

[TK04] G. Taylor and L. Kleeman. Integration of Robust Visual Perception and Control for a Domestic Humanoid Robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 1010–1015, Sendai, Japan, 2004.

[TP91] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Tso05] J. K. Tsotsos. *Neurobiology of Attention*, chapter Computational Foundations for Attentive Processes. Elsevier, 2005.

[VAM05] VAMPIRE. VAMPIRE – Visual Active Memory Processes and Interactive Retrieval, 2005. (IST-2001-34401), http://www.vampire-project.org/ (Date: May 31, 2007).

[VCSS01] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt Visual Attention for a Humanoid Robot. In *Proc. Int. Conf. on Intelligence in Robotics and Autonomous Systems (IROS)*, pages 2332–2337, Hawaii, 2001.

[VGH03] R. T. Vaughan, B. P. Gerkey, and A. Howard. On Device Abstractions For Portable, Reusable Robot Code. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 3, pages 2421–2427, Las Vegas, NV, 2003.

[VJ01] P. Viola and M. Jones. Robust Real-time Object Detection. In *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001.

[W3C06] W3C. Resource Description Framework (RDF), 2006. http://www.w3.org/RDF/ (Date: May 31, 2007).

[Wac01] S. Wachsmuth. *Multi-modal Scene Understanding Using Probabilistic Models*. Dissertation, Bielefeld University, Faculty of Technology, 2001.

[WC05] P. S. P. Wang and C. H. Chen, editors. *Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing Co., 3rd edition, 2005.

[WFBS04] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer. An XML Based Framework for Cognitive Vision Architectures. In *Proc. Int. Conf. on Pattern Recognition*, number 1, pages 757–760, 2004.

[WKG+06] H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. Steil, H. Ritter, and E. Körner. A biologically motivated system for unconstrained online learning of visual objects. In *Proc. Int. Conf. on Artificial Neural Networks*, Athens, Greece, 2006.

[Wor06a] World Wide Web Consortium (W3C®). Extensible Markup Language (XML), 2006. http://www.w3.org/XML/ (Date: May 31, 2007).

[Wor06b] World Wide Web Consortium (W3C®). XML Schema, 2006. http://www.w3.org/XML/Schema (Date: May 31, 2007).

[WR06a] M. Wünstel and T. Röfer. A Probabilistic Approach for Object Recognition in a Real 3-D Office Environment. In Vaclav Skala, editor, *Proc. of 14th Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision*, pages 41–42, Plzen, Czech Republic, 2006.

[WR06b] M. Wünstel and T. Röfer. Feature Based Registration of Range Images in Domestic Environments. In *Computer Vision and Graphics (ICCVG)*, Computational Imaging and Vision, pages 648–654. Springer; Dordrecht, The Netherlands, 2006.

[WWHB05] S. Wachsmuth, S. Wrede, M. Hanheide, and C. Bauckhage. An Active Memory Model for Cognitive Computer Vision Systems. *KI-Journal, Special Issue on Cognitive Systems*, 19(2):25–31, 2005.

List of Publications

T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer.
Human-like person tracking with an anthropomorphic robot. In Proc. IEEE Int. Conf. on Robotics and Automation (ICRA), p. 1286–1292, Orlando, Florida, 2006.

J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, and G. Sagerer.
A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities. In Proc. IEEE Int. Conf. on Robotics and Automation, p. 3419–3425, Barcelona, Spain, 2005.

A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer.
A multi-modal object attention system for a mobile robot. In Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p. 1499–1504, Edmonton, Alberta, Canada, 2005.

S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer.
Human-style interaction with a robot for cooperative learning of scene objects. In Proc. Int. Conf. on Multimodal Interfaces, p. 151–158, Trento, Italy, 2005. ACM Press.

B. Möller, S. Posch, A. Haasch, J. Fritsch, and G. Sagerer.
Interactive object learning for robot companions using mosaic images. In Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p. 371–376, Edmonton, Alberta, Canada, 2005.

I. Toptsis, A. Haasch, S. Hüwel, J. Fritsch, and G. Fink.
Modality integration and dialog management for a robotic assistant. In Proc. European Conf. on Speech Communication and Technology, Lissboa, Portugal, 2005.

A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer.
BIRON - The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, Proc. Int. Workshop on Advances in Service Robotics, p. 27–32, Stuttgart, Germany, 2004. Fraunhofer IRB.

B. Wrede, A. Haasch, N. Hofemann, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, S. Li, I. Toptsis, G. A. Fink, J. Fritsch, and G. Sagerer.
Research issues for designing robot companions: BIRON as a case study. In P. Drews, editor, Proc. IEEE Conf. on Mechatronics & Robotics, Vol. 4, p. 1491–1496, Aachen, Germany, 2004. Eysoldt, Aachen.

Index

Symbols

A

B

C

D

E

F

G

H

O

P

R

S

T

V

W

X