I Would Like Some Food: Anchoring Objects to Semantic Web Information in Human-Robot Dialogue Interactions

Andreas Persson, Silvia Coradeschi, Balasubramanian Rajasekaran, Vamsi Krishna, Amy Loutfi, and Marjan Alirezaie

Center for Applied Autonomous Sensor Systems (AASS)
Dept. of Science and Technology, Örebro University
Örebro, Sweden
{andreas.persson,silvia.coradeschi,amy.loutfi}@oru.se

Abstract. Ubiquitous robotic systems present a number of interesting application areas for socially assistive robots that aim to improve quality of life. In particular the combination of smart home environments and relatively inexpensive robots can be a viable technological solutions for assisting elderly and persons with disability in their own home. Such services require an easy interface like spoken dialogue and the ability to refer to physical objects using semantic terms. This paper presents an implemented system combining a robot and a sensor network deployed in a test apartment in an elderly residence area. The paper focuses on the creation and maintenance (anchoring) of the connection between the semantic information present in the dialogue with perceived physical objects in the home. Semantic knowledge about concepts and their correlations are retrieved from on-line resources and ontologies, e.g. WordNet, and sensor information is provided by cameras distributed in the apartment.

Keywords: anchoring framework, semantic web information, dynamic system, human-robot dialogue, sensor network, smart home environment.

1 Introduction

Socially assistive robots in combination with smart home environments are increasingly considered as a possible solution for providing support to elderly and persons with disabilities [3]. Such ubiquitous robotic systems present a number of interesting application areas that aim at improving quality of life and at allowing people to live independently in their own home longer. These technological solutions can be used to monitor health condition via physiological sensors and activities recognition, to remind about medicines and appointments, to raise alarms, and to assist in everyday tasks like finding objects, detect if objects are misplaced, and guiding in the execution of tasks.

An important facet to many of these applications is the ability to communicate about and interact with objects that are present in the home. A key challenge is therefore to connect the information provided via the dialogue with the sensor

G. Herrmann et al. (Eds.): ICSR 2013, LNAI 8239, pp. 361–370, 2013.

information gathered by the robot and/or a sensor network. The sensor network is able to assist the robot in performing tasks which are otherwise difficult on its own, e.g. localization, while the robot also serves as a useful point of interaction. Studies in human-robot interactions such as [5] have even shown that smart homes with sensor networks are more readily acceptable when a mobile robot is present as an interaction, i.e. receives and relays information to the inhabitants. The challenge to connect semantic information (e.g. general concepts and object names) to the sensor data that refers to the object (e.g. feature descriptors) has been called the anchoring problem. This connection should be first established and then maintained over time. Anchoring has been initially defined in [4] and then extended to consider large knowledge base system like Cyc [6]. The specific challenge of anchoring objects in the context of social robotics in domestic environments, is both the large possibility/variety of objects that need to be anchored and the fact that each object can be referred to in a number of ways in a dialogue. To allow for the possibility to have a dynamic system without the constraints of the dialogue being defined a-priori, open sources of information such as the web can be used. Spoken dialogue is a natural and effortless medium of communication between humans. This together with the evolution of speech related on-line services, makes spoken dialogue a more and more prominent solution for human-robot interaction [7]. It has also been proven that a robot which is capable of interacting in natural language would be more convenient to use, even for users without technical experience [12].

This paper presents an implemented system in a real home environment that can establish a dialogue with a human user about common household objects. The system consists of a mobile robot together with a set of distributed cameras in the home. The robot can accept requests for finding objects via spoken dialogue with a human user and use stored information provided by the cameras about objects present in the environment to answer the requests. In order to create a robustness in the dialogue, ontologies are used to relate concepts employed by the human to the semantic information stored in the anchors that refer to the specific object present in the home. Using on-line ontologies, i.e. WordNet[1], this information is mined as it is needed rather than stored a-priori, as is the case in previously reported work on anchoring [6]. The novelty of this work is twofold. On one hand, the use of ontologies and in particular on-line ontologies is a novel and important contribution to allow for a flexible handling of objects in dialogue and task performance. On the other hand, the integration of the mobile robot in a smart home allows an efficient bottom up approach of anchoring, where all the significant objects in the environment are anchored and continuously and dynamically updated. A request for a specific object is then matched with the stored information. In previous approaches to anchoring, the focus was on a mobile robot that was processing an anchoring request and mainly finding a matching object to the current image and partly to the stored objects images. These approaches were mainly top down and where initiated by a specific request [9]. The new bottom up development presented in

[1] `http://wordnet.rkbexplorer.com/`

this paper has been made possible by the use of complementary static cameras and most importantly by the efficient use of database techniques allowing the storage and easy retrieval of thousands of objects. Section 2 presents a novel anchoring framework suitable for social interaction; Sect. 3 presents a scenario and experimental results while Sect. 4 concludes the paper.

2 Framework for Improved Anchoring via Social Interaction

The anchoring framework, seen in Fig. 1, is a distributed system consisting of several integrated nodes communicating through ROS (Robot Operating System)[2]. Each node has one or more designated task(s) triggered by system events, other nodes, or human users. The overall architecture is divided into three core modules (described in detail in the following sub-sections): 1) *perceptual anchoring module*, 2) *semantic knowledge module*, 3) *human-robot interface module*.

To facilitate persistent storage of information a MongoDB[3] database is used together with an upper generic query interface, which is integrated with both the *perceptual anchoring module* and the *semantic knowledge module*. Such a setup, with a NoSQL database, provides a scalable and dynamic solution suitable for storage of growing collections of both perceptual sensor data and semantic knowledge collected from on-line ontologies.

2.1 Perceptual Anchoring Module

The Perceptual Anchoring Module consists of both a *perceptual system* and an *anchoring system*, shown in Fig. 1 – No. 1. The perceptual system consists of distributed sensors, namely cameras, in the smart home environment, detecting objects in the environment. The output of the perceptual systems are percepts, which are then fed to the anchoring system. A percept is defined to be a structured collection of measurements assumed to originate from the same physical object. The measurements are described by a set of attributes. In our system, percepts are obtained only from visual features that are processed by the images obtained by the cameras. We denote a percept by π and the specific attributes by Φ, specifically, ϕ_1 refer to binary visual features (FREAK [1]), ϕ_2 refer to color for each percept. Each attribute contains a predicate grounding relation, e.g. (white, color, 255 255 255). However, for the visual features, a more complex method to extract the predicate grounding relation, based on existing image databases was used. This method is outside the scope of this paper but a description can be found in [11].

The anchoring system, follows a bottom-up approach described in [9]. The system receives a percept and invokes a matching algorithm in which the percept is compared against all previously stored anchors in the anchoring system. The

[2] `http://www.ros.org/`
[3] `http://www.mongodb.org/`

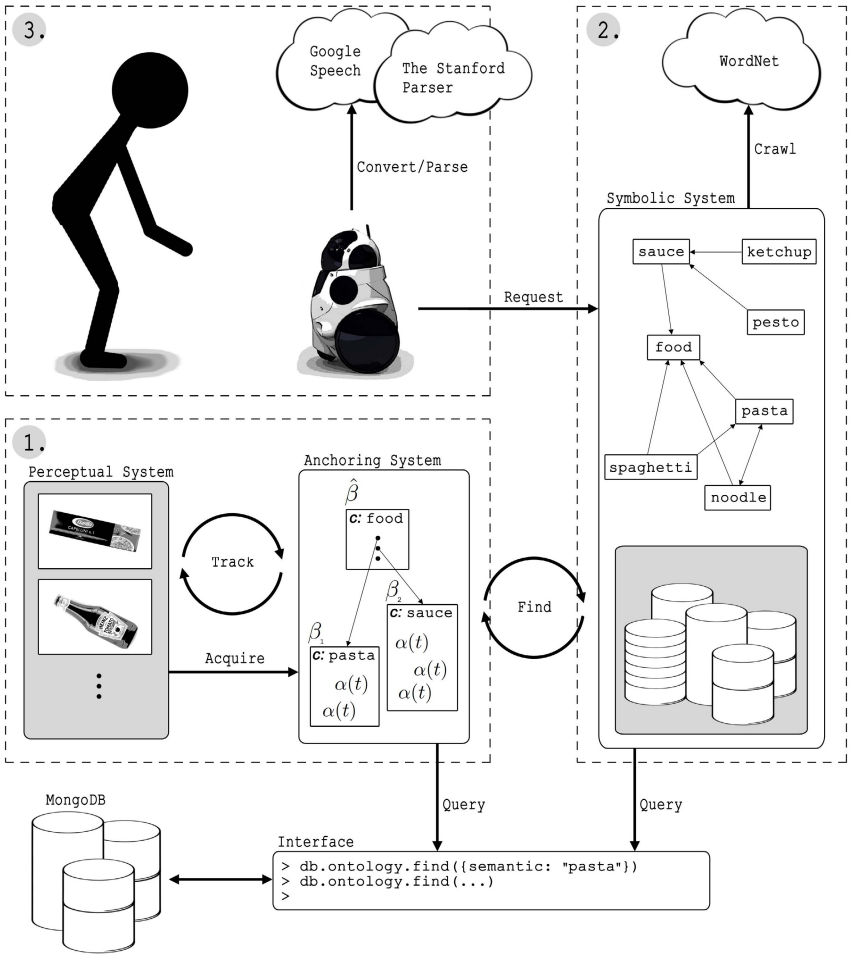

Fig. 1. Overview of the anchoring framework

details of the match algorithm can be found in [11]. Depending on the results of matching, anchors are either created or maintained through two functionalities:

- *Acquire* – creates a new anchor whenever a percept is received which currently does not match a percept of an existing anchor. The new anchor consists of the perceptual description, a symbolic description (which is the set of predicates from the grounding relation), current time t, and a unique identifier.
- *Track* – updates an existing anchor whenever a percept is received which matches the attributes of an existing anchor. Updates the `time` *field* from time $t-k$ to current time t, through the interface towards the database and with the use of a unique identifier of the existing matching anchor.

A third functionality called the *find* functionality can be invoked to find anchors top-down. It is this functionality which is used to find matching anchors requested by the user. This functionality could also provide new, refined or enriched semantic symbolic descriptions as result of a dialogue with a user or as the result of crawling on-line repositories. It is this particular facet which is the novel contribution of this work to the existing work on anchoring.

2.2 Semantic Knowledge Module

The core of this module is a *semantic system*, seen in Fig. 1 – No. 2, consisting of an *ontology* together with a *knowledge base*, both maintained as *collections* in the database. The knowledge base mirrors the result of crawling on-line repositories, while the ontology is stored in the form of hierarchies in a finite search space, where the relation between the various nodes is stored in the form of subsumption relation. A *field* in the *collection* contains a sequential list of all the nodes that have to be traversed in order to search the association between semantic concepts.

As an intermediate layer between the human-robot interface and anchors, this module must be able to receive, interpret and respond upon semantics descriptions about anchors as well as user requests. This is facilitated through three main functionalities:

- *Crawl* – initiates a web crawling in order to update the knowledge base whenever a user is requesting something outside the boundaries of existing knowledge. For this purpose, the WordNet on-line semantic repository is used. This functionality is created in such a way that all the hypernyms and hyponyms/troponyms of a noun/verb are explored recursively till there are no more nodes to explore. The result is then stored in *collections* `nouns`/`verbs` respectively in the database.
- *Request* – receives and parses user requests. Based on requested objects (nouns) and/or activities (verbs), a search for possible candidates in the ontology is conducted. If a match for the request is available, then a reply is sent immediately, otherwise an additional search is conducted for existing knowledge, and where a *crawl* is initiated in case there exists no knowledge about requested verb and/or noun. In the latter case, the crawling with its results is also synchronized with the regularly invoked *find* function so that the ontology is up to date before being re-searched.
- *Find* – the purpose of this is twofold: 1) update the ontology of the *semantic system* based on perceived anchors, 2) update semantics of the *anchor system* based on new knowledge as result of a dialogue with the user. Upon receiving semantic symbols of perceived anchors, the `ontology` *collection* of the database is updated so that the ontology only consists of 'is-a' hierarchies of concepts perceived at current time t. Furthermore, a comparison is made between the existing knowledge and semantic symbols of perceived anchor such that updated knowledge is sent as response to the *anchoring system*.

2.3 Human-Robot Interface Module

A Q.bo Robot[4] and a set of sensors connected via a network are present in the smart home. The robot is a commercial product using the Festival speech synthesis system [2] together with the Julius continuous speech recognition algorithm [8] for interacting with human users. However, the Julius algorithm requires a given grammar, which would limit the speech input, especially if the system is used for human natural conversations. Therefore, the robot's speech-to-text system is instead consisting of an on-line solution using Google Speech[5] together with the Stanford Parser[6] for parsing the text.

The human-robot interface module, seen in Fig. 1 – No. 3, is initiated by a voice command. The human voice is recorded and sent to the Google Speech service for conversion to text. The resulting text is then parsed using Stanford Parser. The result of the parser is further processed by the robot. Here, the robot tries to understand the meaning and context of the users request prior initiating the *semantic system*. Once the robot has confirmed that the user has made a request, the *request* functionality is called. Upon receiving response, results are interpreted, e.g. `there is 'milk' located in the 'kitchen'`, and synthesized as speech through the Festival algorithm.

3 Evaluation

The evaluation was carried out in a smart home used as test apartment (called Ängen), see Fig. 2 – No. 1, which is located in a unique building complex as part of an initiative to provide complete care facilitates for older people in Örebro, both elderly and independent seniors.

As a reference data for the anchoring system was a collection of 6,830 high resolution images (treated as percepts) of common products found in a typical (Swedish) grocery store used. Those images were used for prior training such that attributes were measured and stored in the database together with a semantic symbols for each set of attributes. Where stored attributes were later used for matching new percepts upon arrival at the anchoring system.

3.1 Scenario

The *perceptual system* – consisting of distributed cameras in the smart home environment – registers changes in the environment over a sequential series of frames, captured from one of the cameras located in the kitchen (Fig. 2 – No. 2–3). With the use of computer vision and morphological filters, an area of interest is selected and brought to attention (Fig. 2 – No. 4). Through zooming in on the selected area, five distinct regions of interest corresponding to five percepts are further selected (Fig. 2 – No. 5). Those percepts are passed on

[4] `http://thecorpora.com/`
[5] `http://www.google.com/intl/en/chrome/demos/speech.html`
[6] `http://nlp.stanford.edu/software/lex-parser.shtml`

to the *anchoring system* where specific attributes are measured, namely binary visual features (FREAK [1]) and color. Furthermore, locations of the percepts are given based on the locations of the cameras. The anchoring system performs matching between measured attributes and attributes of all previously stored anchors α_x (as result of prior training), where the system recognize the objects as previously known object and updates the time, t, of existing anchors, results seen in Table 1.

Table 1. Anchored objects perceived by the 'kitchen' camera

UID	Semantic	Color	Location
`ketchup-1`	`ketchup`	`red`	`kitchen`
`milk-1`	`milk`	`white`	`kitchen`
`pasta-1`	`pasta`	`blue`	`kitchen`
`pasta-2`	`pasta`	`red`	`kitchen`
`sauce-1`	`pasta sauce`	`blue`	`kitchen`

At arbitrary time, $t + 1$, the elderly resident of the smart home apartment initiates a conversation with the Q.bo robot in order to ensure him-/herself about that there is something to eat in the apartment (Fig. 2 – No. 6):

- *User:* `Q.bo!`
- *Robot:* `Yes?`
- *User:* `I would like some 'food'.`
- *Robot:* `Let me see what I can find, just a second...`

Once the robot has interpreted the parsed result of the conversation and concluded that the user has made a request (Fig. 2 – No. 7), the *semantic system* is triggered through the *request* function. Upon receiving the request, the *semantic system* searches the 'is-a' hierarchies of the ontology and/or the knowledge base of the concept `food`. In case there is no knowledge about the concept `food`, a search against WordNet is initiated through the *crawl* function. Results of crawling WordNet are further stored as knowledge before the system is synchronized with the *find* function in order to update both the ontology and to enrich the semantic description of perceived anchors based on the new knowledge. In this case, all five perceived objects are possible candidates related to `food` (Fig. 2 – No. 8). However, by looking at similarities between concepts (see following Section 3.2), there exists one semantic candidate corresponding to two objects (`pasta-1` and `pasta-2`) which is more prominent (has higher similarity score) than the others. Information about those objects are sent as response (Fig. 2 – No. 9), where the result is spoken back to the user through the Festival speech synthesis:

- *Robot:* `There are two 'pasta' located in the 'kitchen', one 'red' object and one 'blue' object.`
- *User:* `Thank you Q.bo.`

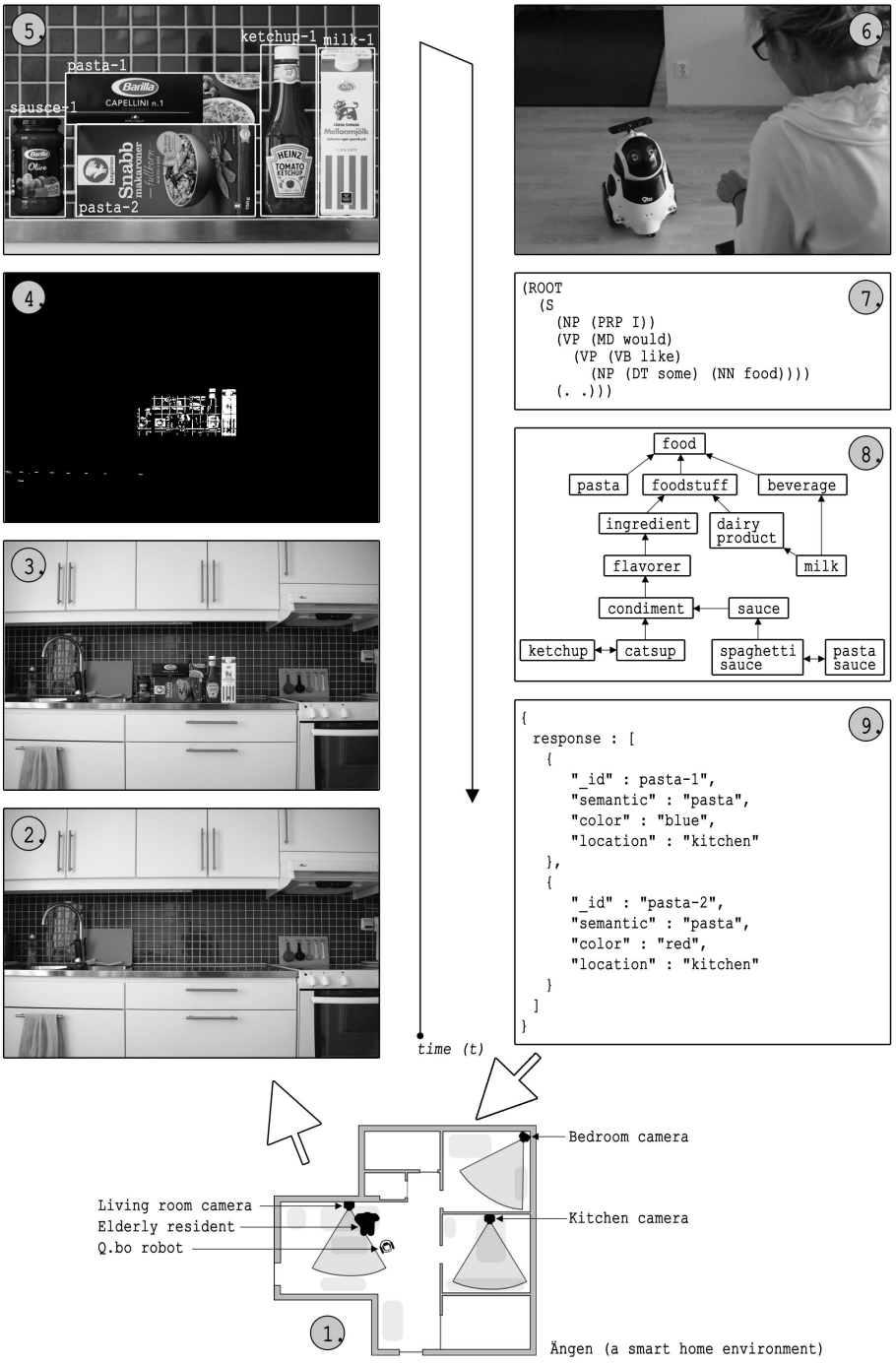

Fig. 2. Overview of test apartment and flow of the steps in evaluation scenario

3.2 WordNet Similarity Search

In a large scale setup, a user query could result in hundreds or thousands of matching candiates, and also ambiguous results, it is of importance to further process matching candiates to find the most prominent candiates. For this purpose was the WordNet::Similarity software package used [10] together with the shortest path algorithm: $y(s_1, s_2) = 1/length(s_1, s_2)$, where s_1 and s_2 are concepts and the *length* between the concpets is the shortest path between the concepts in the 'is-a' hierarchies of WordNet. The most prominent candidates are considered as the ones with highest similarity score y.

As stated previously, the use of on-line repositories, e.g. WordNet, promotes a more dynamic dialogue, and hence, the specific request can vary. The results of a few examples of possible requests like `I would like some 'X'` or `is there something 'X'?`, and the similarity between requested *noun* and known objects is shown in Table 2.

Table 2. Similarity score between a few concepts of example

	s_2 (nouns)				
s_1 (nouns)	`food`	`drink`	`liquid`	`eatable`	`fruit`
`ketchup`	0.1667	0.1429	0.1000	0.1429	0.0667
`milk`	0.3333	0.5000	0.3333	0.2500	0.0833
`pasta`	0.5000	0.2000	0.1667	0.2000	0.0909
`pasta sauce`	0.1429	0.1250	0.0909	0.1250	0.0625

4 Conclusions

The possibility to use technology to contribute to an improved quality of life of elderly and persons with disability and to allow them to live independently in their own homes has rised many expectations and hopes. However the fulfillement of such expectations will only be possible if robust and flexible systems are developed that are suitable for supporting everyday tasks. A key capability of such systems is the possibility to speak about physical objects present in the home using a natural vocabulary. In this paper we have presented a distributed perceptual anchoring framework in a smart home where an ontology can be extended based on requests given by human users and through crawling on-line repositories. This provides for not only an enriched dialogue, but also enhances the semantic symbols anchored to perceptual data in a anchoring system. An example of use of the system has been presented throughout a scenario, where an example of enrichment is shown.

Further developments of this work will consider how to initiate the framework. There exists a certain "chicken or the egg" -situation – percepts are needed in order to match and anchor percepts with semantic symbols, while semantic symbols are needed in the first place in order to initiate an on-line crawling. A solution to this problem is to a-priori train the framework (which was the

case in the scenario presented in this paper), whereas a more complex solution, but also more challenging, would be to involve the user in a dialogue in order to learn objects. User experience and acceptable delays in a natural language dialogue is another area for further evaluation. Delays caused by crawling on-line repositories is highly dependent on the request, recursiveness in the search and bandwidth, and are therefore also difficult to predict.

Acknowledgement. This work has been supported by the Swedish Research Council (Vetenskapsrådet) under the grant number: 2011-6104.

References

1. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: CVPR, pp. 510–517. IEEE (2012)
2. Black, A.W., Taylor, P.A.: The Festival Speech Synthesis System: System documentation. Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK (1997)
3. Coradeschi, S., Cesta, A., Cortellessa, G., Coraci, L., Gonzalez, J., Karlsson, L., Furfari, F., Loutfi, A., Orlandini, A., Palumbo, F., Pecora, F., von Rump, S., Štimec, A., Ullberg, J., Östlund, B.: Giraffplus: Combining social interaction and long term monitoring for promoting independent living. In: Proc. of the 6th Int. Conference on Human System Interaction (HSI 2013), Sopot, Poland (2013)
4. Coradeschi, S., Saffiotti, A.: Anchoring symbols to sensor data: preliminary report. In: Proc. of the 17th AAAI Conf., pp. 129–135. AAAI Press, Menlo Park (2000)
5. Cortellessa, G., Loutfi, A., Pecora, F.: An on-going evaluation of domestic robots. In: Robotic Helpers: User Interaction, Interfaces and Companions in Assistive and Therapy Robotics, pp. 87–91 (2008)
6. Daoutis, M., Coradeschi, S., Loutfi, A.: Cooperative knowledge based perceptual anchoring. International Journal on Artificial Intelligence Tools 21(3) (2012)
7. Kanda, T., Ishiguro, H., Ono, T., Imai, M., Mase, K.: Multi-robot cooperation for human-robot communication. In: IEEE Int. Workshop on Robot and Human Communication (ROMAN 2002), pp. 271–276 (2002)
8. Lee, A., Kawahara, T.: Recent development of open-source speech recognition engine julius. In: Proc. of APSIPA ASC, pp. 131–137 (2009)
9. Loutfi, A., Coradeschi, S., Saffiotti, A.: Maintaining coherent perceptual information using anchoring. In: Proc. of the 19th IJCAI Conf., Edinburgh, UK, pp. 1477–1482 (2005)
10. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity - measuring the relatedness of concepts. In: HLT-NAACL 2004: Demonstration Papers, May 2-7, pp. 38–41. Association for Computational Linguistics, Boston (2004)
11. Persson, A., Loutfi, A.: A hash table approach for large scale perceptual anchoring. In: Proc. of IEEE Int. Conference on Systems, Man and Cybernetics (SMC 2013), Manchester, UK (2013)
12. Spiliotopoulos, D., Androutsopoulos, I., Spyropoulos, C.D.: Human-robot interaction based on spoken natural language dialogue. In: Proc. of the European Workshop on Service and Humanoid Robots (ServiceRob 2001), Santorini, Greece, pp. 25–27 (2001)