The Golem Team, RoboCup@Home 2014

Team Leader: Luis A. Pineda[1], Caleb Rascon, Gibran Fuentes, Varinia Estrada, Arturo Rodriguez, Ivan Meza, Hernando Ortega, Mauricio Reyes, Mario Peña, Joel Duran, Erandu Campos, Sebastian Chimal, and Albert Orozco

[1] lpineda@unam.mx

http://turing.iimas.unam.mx/~luis

Computer Science Department
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)
http://golem.iimas.unam.mx

Abstract. In this work we describe the Golem Team and the latest version of the robot Golem-II+. This is the fourth time the Golem team participates on the RoboCup@Home Competition. The design of our robot is based on a conceptual framework that is centered on the notion of dialogue models with the interaction-oriented cognitive architecture (IOCA) and its associated programming environment, SitLog. This framework provides flexibility and abstraction for task description and implementation, as well as a high level modularity. The tasks of the RoboCup@Home competition are implemented under this framework using a library of basic behaviours.

1 Team Members

Robot: Golem-II+.
Academics:

Dr. Luis A. Pineda. Dialogue Management, SitLog, Task Structure and Robotic Behavior, Knowledge Representation and Cognitive Architecture.
Dr. Caleb Rascon. Audio, Sound Localization and Navigation.
Dr. Gibran Fuentes. Vision and Object Manipulation.
Ms. Varinia Estrada. Visual modelling and Technical Support.
Dr. Ivan V. Meza. Speech Recognition and Language.
M.Sc. Hernando Ortega. Electro-mechanical Devices.
M.Sc. Mauricio Reyes Castillo. Industrial Design.
Dr. Mario Peña. Electronics and Instrumentation.
Mr. Joel Duran. Electronics and Instrumentation.

Students:

M.Sc. Arturo Rodriguez-Garcia. Person Tracking and SitLog Behaviours.
Erandu Campus Technical Support.
Sebastian Chimal Technical Support.
Albert Orozco Technical Support.

2 Group Background

The Golem Group is a research group focused on robotics mainly on the cognitive modeling of the interaction between humans and robots. The group was created within the context of the project "Diálogos Inteligentes Multimodales en Español" (DIME, Intelligent Multimodal Dialogues in Spanish) in 1998 at IIMAS, UNAM where it has been established since. The goals of the DIME project were the analysis of multimodal task-oriented human dialogues, the development of a Spanish grammar, speech recognition in Spanish, and the integration of a software platform for the construction of interactive systems with spoken Spanish. By 2001 the group started the Golem project with the purpose of generalizing the theory for the construction of intelligent mobile agents, in particular the Golem robot. A first result was a version of a theory for the specification and interpretation of dialogue models which is still a corner stone in the group's philosophy [11].

Several versions of the Golem robot were demonstrated at the Universum Science Museum in which Golem interacted with visitors. In 2002, the robot had a simple conversation and followed movements commands. In 2006, it guided a poster session. Finally, in 2009 we presented the module "Guess the card: Golem in Universum" in which children played a game [9].

In 2010, we started the development of Golem-II+ our current service robot. Golem-II+ incorporates an innovative explicit cognitive architecture that, in conjunction with the dialogue model theory and program interpreter, constitutes the theoretical core of our approach [12,15].

Since 2011, we have participated at the RoboCup@Home competition: Istanbul 2011, Mexico 2012 and Netherlands 2013. We have also participated on the local Mexican competitions in 2012 (1st place) and 2013, and German Open in 2012 (3rd place). All of which provided important feedback for the robot's performance. In particular, at the RoboCup@Home 2013 the team was awarded the Innovation Award of the league for our demo in which the robot uses its audio-localization system to perform a waiter role in a noisy enviroment. This demo incorporated the capability of spatial reasoning to the navigation subsystem, including the ability of facing the interlocutor during human-robot interactions [17] and provided the possibility for the robot to interact with more than one agent at a time [19].

The current version of our robot expands our previous developments [13]. This version uses a new set of modular behaviours programmed in SitLog[14], a new knowledge base system, a new system for detecting and tracking heads at the distance, and a new audio-activity tracker. In terms of hardware, we have added a new set of cameras to be used by the the computer vision module, and two new grippers with more preasure range for the arms.

3 An Interaction-Oriented Cognitive Architecture

The behavior of our robot Golem-II+ is regulated by an Interaction Oriented Cognitive Architecture (IOCA) [12,15]. The IOCA architecture specifies the types of modules which integrate our system. A diagram of IOCA can be seen in Figure 1.

Recognition modules encode external stimuly into specific modalities (e.g., speech into utterances transcriptions, images to SIFT features). *Interpreter* modules assign a meaning to those messages from different modalities (e.g., from utterances or SIFT features to a semantic representation, $name(golem)$ or $object(juice)$). On the other side, *Specification* modules specify global parameters into particular ones for the actions (e.g., *kitchen* the x,y points). *Render* modules are in charge to execute the actions (e..g, perform navigation actions to arrive to the kitchen). In the case of the dialogue manager module there is only one of its type. This is in charge to manage the execution

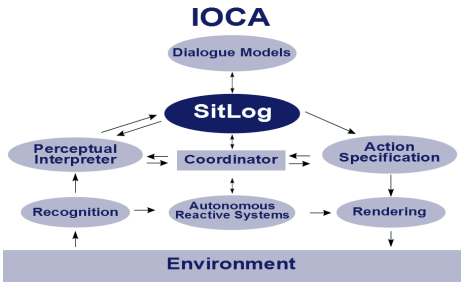

Fig. 1. Interaction Oriented Cognitive Architecture (IOCA).

of the task. A more detailed explanation is presented in section 4.1. Communication between the modules is done using the OAA libraries [1].

Reactive behavior is reached by tightly joining recognition and render modules into *Autonomous Reactive Systems* (ARSs). Example of these system are the Autonomous Navigation System (ANS) and the Autonomous Position and Orientation Source of Sound Detection System (APOS) to allow the robot to face its interlocutor reactively.

4 Software

We organize our software in modules for different skills and the IOCA architecture manages the connectivity among these modules. Additionally, for some of these skills we have abstracted some basic behaviours which are programmed in SitLog. These behaviours encapsulate skills and their recovery strategies. Example of such behaviours are to detect a person, to see an object or to visit a sequence of points. Next we present the main modules and its associated behaviours.

4.1 Dialogue Manager

The central communication and control structure of Golem is defined through modular schematic protocols that we called Dialogue Models (DM). DMs represent the task structure of application domains. DMs are specified in SitLog (Situations and Logic) [14][1], a declarative programming language developed within the context of the project. DMs have a diagrammatic representation as Recursive Transition Networks, where nodes represent world situations and edges represent expectations and actions pairs, and situations can stand for fully embedded tasks. SitLog has also an embedded functional language for the declarative specification of control and content information. Expectations, actions and situations are specified through basic expressions of this language or through functions that are evaluated dynamically, supporting large abstraction in this dimension too. SitLog's interpreter is coded in Prolog and the specification of DMs follows closely the Prolog's notation. SitLog's interpreter is the central component of the IOCA architecture, and evaluates DMs continuously during the execution of the tasks, and also coordinates reactive and deliberative behavior.

[1] http://golem.iimas.unam.mx/sitlog

4.2 Knowledge Representation

Golem has a central knowledge representation system consisting of a KB manager with its knowledge repository and administration procedures. Knowledge is specified as a class taxonomy with inheritance and supports naturally the expression of defaults and exceptions. The system permits the expression of properties of classes and relations between classes and also the expression of individuals of each class with their particular properties and relations too. Conflicts between particular and general properties and relations are handled through the criterium of specificity, such that properties and relations of individuals have precedence over the properties and relation of their classes. All objects within the KB can be updated dynamically and the scheme behaves non-monotonically. The KB system is coded in Prolog and the KB-services can be used within the body of DMs directly. The KB system has been fully design and developed within the context of the project.

4.3 Vision

Vision is carried out via various vision modules, described hereafter

Face and Head Detection, Tracking and Recognition. OpenCV functions are used to perform face and head detection, tracking and recognition. Face detection is carried out by using the Viola-Jones Method [21]. Face recognition is based on Eigenfaces technique [20]. For the head detection we use a Histogram of Oriented Gradients (HOG) models that were created in house[3]. Tracking of face and head is carried out via a technique based on data association with the Hungarian Algorithm and Kalman Filtering [6].

We had found that our HOG head detector is more reliable to detect persons at distance, so we have devised a behavior for the detection of persons based on alternating face and head detection in order to enhance the performance. Similar behaviours have been created for the recognition and memorization of a user. During the search, we take advantage of our 2-DOF neck movement capability to enhance the search.

Object Recognition. This capability is performed using the MOPED framework proposed by Collet et. al. [2]. During the training stage, several images are acquired from each object that is to be recognized, and SIFT features [7] are extracted from each image. Structure from Motion is applied to arrange all the features from all the images and obtain a 3D model of each object. In the recognition phase, the SIFT features are obtained from the visual scene and, with 3D model at hand, hypotheses are made in an iterative manner using Iterative Clustering Estimation [2]. Finally, Projective Clustering [2] is used to group the similar hypothesis and provide one for each of the objects being observed. A behaviour associated to this capability is in charge of returning the position and orientation of a given object, or if not specified, it can return the closest one in the scene.

Person Tracking. For person identification, we use the Person Tracker that is part of the OpenNI driver, where several blobs in the visual scene are hypothesized as being a person. A person that performs a certain known action is then labeled in the visual scene as the one to be tracked.

Plane detection. We use the Point Cloud Library to detect planes. In particular, we focus on horizontal planes which could correspond to tables. We use this information to put safely an object on the tab table.

4.4 Arm and Neck Manipulation

The 3-DOF robotic arms were built in-house. These are mounted on the robot on its mobile platform (modified in-house such that is its height can be controlled via software), providing a fourth DOF. The robotic arms are based on Robotis Dynamixel motors for movement, and are controlled via a Servo Controller, which, in turn, accepts commands of the type of "park", "grasp object", and "offer object". The latter two are able to accept distance and angle arguments.

When it comes to the "grasp" command, the hands of the robotic arms come equipped with three IR sensors which are employed for a local object search, having arrived to the desired angle and distance.

Associated to the arm there are the *take* and *deliver* behaviours. The *take* behaviour will grasp an object taking into account the position and orientation information obtained by the object recognition module. Additionaly, it positions the robot so the object is in reach. The *deliver* behaviour is in charge to deliver an object; it can do it to a specific position, or put it in a table using the plane detection module.

The 2-DOF robotic neck was also built in-house, and it is mounted over the upper base of the robot. This neck allows the range of the Kinect and the color camera to be shifted vertically and horizontally providing a wide area of recognition. In addition, a directional microphone is mounted over the horizontal DOF for the same purpose.

4.5 Speech Recognition and Synthesis

We use a robust live continuous speech recognizer based on the PocketSphinx software, coupled with the Walt Street Journal (WSJ) acoustic models. For the language models, we hand-crafted a corpus for each of the tasks, and made the ASR be able to switch from one to the other, depending on the context of the dialogue (A yes/no language model for confirmation, a name language model for when the user is being asked their name, etc.). This module is a Recognizer in the IOCA framework [8].

Similarly, for the speech synthesis we use open tools, in particular the Festival TTS package. From the point of view of the IOCA framework, this is a Rendering module.

Both recognition and synthesis are as an autonomous system so that the robot does not speak while listening or viceversa.

4.6 Language Interpretation

In this version of the system there are two strategies for the language interpretation. The first one is a shallow semantic strategy which uses word and expression spotting. For this strategy, regular expressions and their meanings are stored. The natural language interpreter tries to match the regular expressions to the orthographic transcriptions that are similar to the expectations of the system. The second strategy is a deep semantic parser based on the GF formalism [16]: several grammars were defined specifically for fine grained semantics tasks, such as General Purpose Service Robot.

4.7 Audio

The GPL software JACK is used to create an all-encompassing simulated sound card that can be accessed by different audio clients at the same time. Two audio clients were created as Recognition Modules in IOCA.

Audio-Localization. This module provides a robust direction-of-arrival estimation in near-real-time manner in mid-level reverberant environments, throughout the 360° range. The signals from the three microphones are set in an equilateral triangle, which provide three measured delay-comparisons. This provides redundancy to the direction-of-arrival estimation, as well as a close-to-linear mapping between delay measurements and direction-of-arrival estimations [17]. This module, in conjunction with the Reactive Navigation Module (described later), compose the Autonomous Position and Orientation Source of Sound Detection System (APOS). In addition, a multi-DOA estimation is employed if there are more than one user in the environment [18,19].

4.8 Navigation

The Autonomous Navigation System (ANS) inside IOCA is divided in two parts: The Semantic Planner and the Reactive Navigation Module (RNM). The former interacts directly with the Dialogue Manager, which only needs to provide a pre-specified label of a location of where it is desired to move. The label is realized into a set of (x, y) Cartesian coordinates and the route that the robot will partake is deduced via the Dijkstra Algorithm [4], defined by a series of intermediate points. These are obtained from a topological map created from a weighted adjacency matrix overlaid over the map of the area. Between each intermediate point, the RNM takes over by carrying out obstacle evasion, via, in conjunction, Nearness Diagram [10] (by default) and Smooth Nearness Diagram [5] (as fall-back). In addition, this module provides simple moving capabilities (turn θ degrees, move Z meters to the front, etc.), using the Semantic Planner as a proxy to the Dialogue Manager.

The navigation module has associated one behaviour which coordinates the movement of the robot in its different versions: relative or absolute, topological places or coordinates.

4.9 Software Libraries

Both the robot internal computer and the external laptop run the Ubuntu 12.04 operating system. Table 1 shows which software libraries are used by the IOCA modules and Golem-II+'s hardware.

5 Description of the Hardware

The "Golem-II+" robot (See Fig. 2) will be used, which is composed by the following hardware:

- PeopleBot™ robot (Mobile Robots Inc.)
  - One protective 5-bumper arrays.
  - Speakers.
  - Internal computer VersaLogic EBX-12.
- Dell Precision M4600 laptop computer.
- Point Grey Flea USB 3 camera
- Microsoft Kinect Camera
- Hokuyo UTM-30LX Laser
- Shure Base Omnidirectional microphones x3
- RODE VideoMic directional microphone
- M-Audio Fast Track Ultra external sound interface
- Infinity 3.5-Inch Two-Way loudspeakers x2
- In-house robotic arms, gripper and neck

Table 1. Software Libraries used by the IOCA Modules and Hardware of Golem-II+

Module	**Hardware**	**Software Libraries**
Dialogue manager	–	SitLog, Sicstus Prolog
Knowledge-base	–	Prolog
Vision	Kinect, WebCam, In-house Built Neck	SVS, OpenCV, PCL, and OpenNI libraries
Voice recognition	Directional Microphone, External Sound Card	JACK, PocketSphinx
Voice synthetizer	Speakers	PulseAudio, Festival TTS
Navigation	Bumpers, Laser, Odometric Sensors	Player
Object Manipulation	In-house Built Robotic Arm	Dynamixel RoboPlus
Camera/Mic. Movement	In-house Built Robotic Neck	Dynamixel RoboPlus

Fig. 2. The Golem-II+ robot.

Acknowledgments

Golem-II+ was financed by CONACyT project 178673, by PAPIIT-UNAM project IN107513 and SECITI project ICyTDF/209/2012. We would like to thank the students that have provided us support outside the competition:

- **Rogelio Romero.** Face tracking development.
- **Alex Crespo.** Support on testing.

References

1. Cheyer, A., Martin, D.: The Open Agent Architecture. Journal of Autonomous Agents and Multi-Agent Systems 4(1), 143–148 (March 2001), http://www.ai.sri.com/~oaa
2. Collet, A., Martinez, M., Srinivasa, S.S.: The MOPED framework: Object Recognition and Pose Estimation for Manipulation. The International Journal of Robotics Research 30, 1284–1306 (2011)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893 vol. 1 (2005)

4. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959), http://dx.doi.org/10.1007/BF01386390, 10.1007/BF01386390
5. Durham, J., Bullo, F.: Smooth Nearness-Diagram Navigation. In: Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on. pp. 690–695 (sept 2008)
6. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME–Journal of Basic Engineering 82(Series D), 35–45 (1960)
7. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004)
8. Meza, I., Rascon, C., Pineda, L.: Practical Speech Recognition for Contextualized Service Robots, Lecture Notes in Computer Science, vol. 8266, pp. 423–434. Springer Berling Heidelberg (2013)
9. Meza, I.V., Salinas, L., Venegas, E., Castellanos-Vargas, H., Alvarado-González, M., Chavarría-Amezcua, A., Pineda, L.A.: Specification and Evaluation of a Spanish Conversational System Using Dialogue Models. Advances in Artificial Intelligence - IBERAMIA 2010 6433 (2010)
10. Minguez, J., Member, A., Montano, L.: Nearness Diagram (ND) Navigation: Collision Avoidance in Troublesome Scenarios. IEEE Transactions on Robotics and Automation 20, 2004 (2004)
11. Pineda, L.A.: Specification and Interpretation of Multimodal Dialogue Models for Human-Robot Interaction. In: Sidorov, G. (ed.) Artificial Intelligence for Humans: Service Robots and Social Modeling, pp. 33–50. SMIA, Mexico (2008)
12. Pineda, L.A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., Villaseñor, L.: The Corpus DIMEx100: Transcription and Evaluation. Language Resources and Evaluation 44, 347–370 (2010)
13. Pineda, L., Group, G.: The Golem Team, RoboCup@Home 2013. In: Proceedings of RoboCup 2013 (2013)
14. Pineda, L., Salinas, L., Meza, I., Rascon, C., Fuentes, G.: SitLog: A Programming Language for Service Robot Tasks. International Journal of Advanced Robotic Systems 10(538) (2013)
15. Pineda, L.A., amd Héctor H. Avilés, I.V.M., Gershenson, C., Rascón, C., Alvarado, M., Salinas, L.: IOCA: An Interaction-Oriented Cognitive Architecture. Research in Computer Science 54, 273–284 (2011)
16. Ranta, A.: Grammatical Framework: Programming with Multilingual Grammars. CSLI Publications, Stanford p. 340 (2011), iSBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth)
17. Rascón, C., Avilés, H., Pineda, L.A.: Robotic Orientation towards Speaker for Human-Robot Interaction. Advances in Artificial Intelligence - IBERAMIA 2010 6433, 10–19 (2010)
18. Rascon, C., Pineda, L.: Lightweight Multi-DOA Estimation on a Mobile Robotic Platform. In: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2012, WCECS 2012, 24-26 October, 2012, San Francisco, USA. pp. 665–670 (2012)
19. Rascon, C., Pineda, L.: Multiple Direction-of-Arrival Estimation for a Mobile Robotic Platform with Small Hardware Setup, Lecture Notes in Electrical Engineering, vol. 247, pp. 209–223. Springer Netherlands (2014)
20. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
21. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. I–511–I–518 (2001)