The Golem Team, RoboCup@Home 2017

Team Leader: Luis A. Pineda
Caleb Rascon, Gibran Fuentes, Arturo Rodríguez, Hernando Ortega, Mauricio Reyes, Noé Hernández, Ricardo Cruz, Ivette Vélez, and Marco Ramírez

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-III. This is the fifth 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 behaviors. In addition, over this framework, a system that carries out diagnostics, decision making and planning has been developed as an inference machine platform with which error detection and recovery is possible.

1 Team Members

Robot: Golem-III.
Academics:

- **Dr. Luis A. Pineda.** SitLog, Knowledge Representation, Inference and Cognitive Architecture.
- **Dr. Caleb Rascon.** Robot Audition, Navigation and Module Integration.
- **Dr. Gibran Fuentes.** Vision and Object Manipulation.
- **M.Sc. Hernando Ortega.** Robotic Platform Development and Embedded Software Control.
- **M.Sc. Mauricio Reyes Castillo.** Industrial Design and Emotion Expression.
- **M.Sc. Noé Hernández.** Object Modeling and SitLog Behaviors.

Students:

- **M.Sc. Arturo Rodríguez Garcia.** Person Recognition and Tracking, SitLog Behaviors, and Knowledge Base and Inference Programming.
- **M.Sc. Ricardo Cruz.** Object Modeling and Health-care Applications.
- **B.Sc. Ivette Vélez.** SitLog Behaviors and Robot Audition.
- **B.Sc. Marco Ramírez.** Point Cloud Processing and Plane Detection.

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 [7].

Since 2011, we have participated at the RoboCup@Home competition: Istanbul 2011, Mexico 2012, Netherlands 2013 and Germany 2016. We have also participated on the local Mexican competitions in 2012 (1st place), 2013 and 2016 (2nd place), 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.

During the span of 2014 and 2015, the team developed the iteration of the Golem presented in Germany 2016, Golem-III. This version uses a set of modular behaviors programmed in SitLog[9], a knowledge base system[11], a system for detecting, identifying and tracking persons, and a audio-activity tracker. In terms of hardware, this implementation uses a robotic torso, which includes a 2-DOF robotic neck and two 5-DOF robotic arms. In the iteration presented here, Golem-III has a new platform for diagnostic, decision making and planning, which was built as an inference machine platform that is useful for error detection and recovery.

3 An Interaction-Oriented Cognitive Architecture

The behavior of our robot Golem-III is regulated by an Interaction Oriented Cognitive Architecture (IOCA) [8,10]. 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 stimuli 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). 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

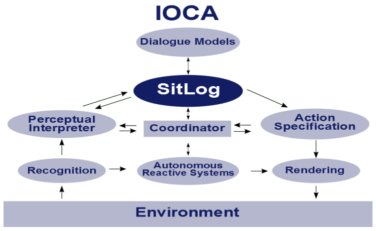

Fig. 1. Interaction Oriented Cognitive Architecture (IOCA).

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 of the task. A more detailed explanation is presented in section 4.1.

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. In this section, we present these modules and their associated behaviors.

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 structure of a given task. DMs are specified in SitLog (Situations and Logic) [9][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.php

4.2 Knowledge Representation

Golem has a central knowledge representation system[11] 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, relations between classes, and the expression of individuals of each class with their particular properties and relations. Conflicts between particular and general properties and relations are handled through the criteria 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. It has been fully design and developed within the context of the project.

4.3 Diagnostic, Decision Making and Planning

Based on the Knowledge Representation, a diagnostic system has been implemented over the represented world in a symbolic fashion. From these diagnostics, error detection can be carried out, such as when an object is expected to be in a certain location but it is not. Over these diagnostic system, decision making and planning can be carried out via an inference machine platform. This platform has been developed over which different kinds of inference techniques can be built, and obtain information such as establishing why the object is not where it is expected. Based on this, a plan can then be devised for error recovery, such as proposing a new location where the object may be.

4.4 Vision

Vision is carried out via various vision modules, described hereafter.

Face and Head Recognition. Face detection and recognition are carried out via the Viola-Jones Method [16] and Eigenfaces [15]. For head detection we use Histogram of Oriented Gradients (HOG) models created in house [3]. During the search, we take advantage of our 2-DOF neck movement capability to enhance the search.

Object Recognition. Objects with textures are recognized using the MOPED framework proposed by Collet et. al. [2]. SIFT-based object models are obtained using Structure from Motion. During recognition, SIFT features [6] are obtained and object hypotheses are made in an iterative manner using Iterative Clustering Estimation and grouped by Projective Clustering [2].

Objects without textures are localized using the Point Cloud Library (PCL). First, horizontal planes are detected to correspond with a table. Then, PCL is used to localize objects using the points that emerge from such planes in the point cloud.

Person Tracking, Gesture Estimation and Soft Biometric Identification. Kinect 2 SDK is used to detect persons and their respective skeletons. This information is used to estimate if persons are waving or pointing with their hands and to classify if they are standing, sitting or laying on the floor. The orientation of persons in relation to the robot is also estimated to determine if they are facing the robot. The skeleton is used to learn and identify persons based on their clothes. Different views of the same person indexed by their orientation angle are stored in the soft biometric database. The identification by clothes is intended to be used in situations where face recognition is not suitable. The current angle of the user to be identified is used to select the nearest view of each person stored in the soft biometric database, and then comparing small patches semantically tagged, extracted from different parts of the body (arms, chest, legs, etc.). Microsoft Cognitive Services is used to obtain description data from the person, such as gender, age, if the person is wearing glasses and facial hair.

A laser installed at the height of the chest of the robot is used to carry out valley-location technique to propose as the direction-distance pair of the person location.

4.5 Arm and Neck Manipulation

The 5-DOF robotic arms were built in-house and are controlled via a Servo Controller. These are mounted on the robot on its torso, the height of which can be controlled via two electronic pistons, providing a sixth DOF. The central upper part of the torso, a seventh DOF is provided for both arms, which acts as a clavicle that extends the length of the manipulation range.

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.

4.6 Speech Recognition and Synthesis

Based on the Windows Speech API, the ASR is able to switch between language models 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.). The ASR is kept idle until a recognition is requested by the Dialogue Manager. In addition, the speech synthesis is also based on Windows Speech API, using the US Male voice. Both recognition and synthesis are an autonomous system so that the robot does not speak while listening or vice-versa.

4.7 Language Interpretation

In this version of the system, the language interpretation is based on a parser implemented in Prolog using Definite Clause Grammars, mounted over of a treebased structure for re-usability. All rules and terminals are objects stored in the knowledge base.

4.8 Audio Localization

It provides a robust direction-of-arrival estimation in mid-reverberant environments, throughout the 360° azimuth range, from a 3-microphone array [1] via a redundant direction-of-arrival estimation [13]. In addition, a multi-DOA estimation is employed if there are more than one user in the environment [14].

4.9 Navigation

It is based on the ROS Navigation Stack that uses GMapping for map creation [5] and Adaptive Monte-Carlo Localization (AMCL) [4]. We also implemented a Semantic Proxy that carries out topological translation between a label of a custom location and its coordinates and robotic pose. The Navigation system can provide several versions of movement: relative or absolute, topological places or coordinates, normal or fine movement, using a pre-made map or carry out automatic mapping.

4.10 Software Libraries

Both the robot internal computer and the external laptop run the Ubuntu 16.04 operating system, and inter-modular communication is done using ROS Kinetic [12]. Table 1 shows which software libraries are used by the IOCA modules and Golem-III's hardware.

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

Module	**Hardware**	**Software Libraries**
Dialogue manager	–	SitLog, SWI Prolog
Knowledge-base	–	SWI Prolog
Vision	Kinect 2 and Flea3 Camera	MOPED, PCL, Kinect 2 SDK, MS Cognitive Services
Robot Audition	Rode VideoMic, 8SoundsUSB Sound Card	Windows Speech API, JACK
Voice synthetizer	Speakers	Windows Speech API
Navigation	Lasers, Odometric Sensors	ROS Navigation Stack
Object Manipulation	Custom Robotic Torso	Dynamixel RoboPlus
Camera/Mic. Movement	Custom Robotic Neck	Dynamixel RoboPlus

5 Description of the Hardware

The "Golem-III" robot (See Fig. 2) is composed by the following hardware:

- PatrolBot^{TM} robot
  - 8-sensor sonar array
  - Two 5-bumper protective arrays
  - Infinity 3.5-Inch two-way loudspeakers
  - On-board computer Cobra EBX-12
  - Sick LMS-500 Laser
- Two Dell Precision M7510 laptop computers
- Hokuyo SOKUIKI laser range finder
- Black Box 5-port usb-powered ethernet switch
- Microsoft Kinect 2 camera
- Point Grey Flea USB 3 high-resolution camera
- 8SoundsUSB audio interface
- 3 miniature microphones
- RODE VideoMic directional microphone
- In-house robotic torso, arms and neck

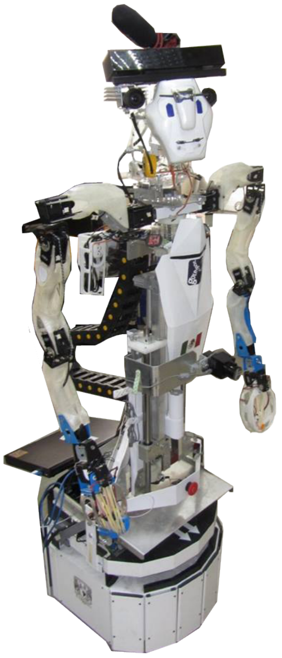

Fig. 2. The Golem-III robot.

Acknowledgments

Golem-III was financed by CONACyT project 178673, by PAPIIT-UNAM project IN107513 and SECITI project ICyTDF/209/2012. We would also like to thank the following personnel for their support outside of the competition:

- Uriel Ortiz. CAD modeling and manipulation development.
- Dennis Mendoza. Laser acquisition, filtering, and analysis.

References

1. Abran-Cote, D., Bandou, M., Beland, A., Cayer, G., Choquette, S., Gosselin, F., Robitaille, F., Kizito, D.T., Grondin, F., Letourneau, D.: USB Synchronous Multichannel Audio Acquisition System
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. Dellaert, F., Fox, D., Burgard, W., Thrun, S.: Monte Carlo localization for mobile robots. In: Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on. vol. 2, pp. 1322–1328 vol.2 (1999)
5. Grisetti, G., Stachniss, C., Burgard, W.: Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters. Robotics, IEEE Transactions on 23(1), 34–46 (Feb 2007)
6. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004)
7. 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)
8. 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)
9. 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)
10. Pineda, L.A., Meza, I.V., Avilés, H., Gershenson, C., Rascon, C., Alvarado-González, M., Salinas, L.: IOCA: Interaction-Oriented Cognitive Architecture. Research in Computer Science 54, 273–284 (2011)
11. Pineda, L.A., Rodríguez, A., Fuentes, G., Rascón, C., Meza, I.: A light non-monotonic knowledge-base for service robots. Intelligent Service Robotics pp. 1–13 (2017)
12. Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: ROS: an open-source Robot Operating System. In: ICRA Workshop on Open Source Software (2009)
13. 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)
14. Rascon, C., Fuentes, G., Meza, I.: Lightweight multi-DOA tracking of mobile speech sources. EURASIP Journal on Audio, Speech, and Music Processing 2015(11) (2015)
15. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
16. 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)