Integrating Graph-Based Vision Perception to Spoken Conversation in Human-Robot Interaction

Wendy Aguilar and Luis A. Pineda

Departamento de Ciencias de la Computación,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Apdo. Postal 20-726, México, D.F., 01000

Abstract. In this paper we present the integration of graph-based visual perception to spoken conversation in human-robot interaction. The proposed architecture has a dialogue manager as the central component for the multimodal interaction, which directs the robot's behavior in terms of the intentions and actions associated to the conversational situations. We tested this ideas on a mobile robot programmed to act as a visitor's guide to our department of computer science.

1 Introduction

Human-robot interaction can be modeled in terms of flexible protocols focused on the expression and interpretation of intentions in interactive situations, and the execution of actions that satisfy such intentions. Such protocols can involve a range of modalities, like spoken language, pointing actions and vision for the input, for instance, and spoken language, the display of pictures and motor actions of physical devices (like mobile robots) for output. Conversational protocols may range from the very rigid and deterministic schemes involved in menu-based interaction to the rich and flexible patterns of natural language, like the ones exhibited in the so-called practical dialogues [1]. This work presents an integration of a graph-based vision perception module into a service robot with multimodal capabilities, that is based on the specification and interpretation of multimodal dialogue models, as presented in [2].

The proposed architecture was tested on a mobile robot, called Golem, that acts as the guide of a poster session about the research projects in our department of computer science. During the visit, the robot can interact with the user through its visual capabilities and recognize posters within the context of a spoken conversation in Spanish. Once a poster is recognized, the robot is able of explain it in spoken Spanish, with the support of other modalities like texts and graphics which are displayed on a screen. Similar multimodal vision and language projects are presented in [3], [4] and [5]; however, none of these include our notion of dialogue model specification and interpretation and use graph based algorithms for vision recognition.

J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 789–796, 2009.

The paper is organized as follows. Section 2 presents the proposed architecture and explains briefly each module, making emphasis on the integration and implementation of the vision module. Section 3 presents tests and results for the object recognition module and the system as a whole. Finally, the conclusions are presented in section 4.

2 Architecture

In this paper we propose to integrate vision to spoken conversation into a three level agent architecture where vision and language are interpreted through analogous processes (see Fig. 1(a)). The lower level is devoted to the recognition of modality specific information; the intermediate level assigns an interpretation to the images, either visual or linguistic, recovered by the speech and vision recognition systems, in terms of the expected intentions and the potential actions of the agent in the interpretation situation. In the third level, the multimodal conversational protocols (i.e. a set of conversational situations, with their associate intentions and actions) are represented in a modality independent fashion.

In this architecture perceptual modules have two components: 1) A modality specific recognition agent, and 2) a perceptual interpreter agent, see Fig. 1(b). The first component is responsible of sensing a visual or an acoustic signal and create its corresponding uninterpreted modality specific image. The perceptual interpretation agent proper assigns an interpretation to such image in terms of memory, where images with their corresponding interpretation are stored, and the current expected intentions of the agent, as specified in the dialogue model. The final output of the *Perceptual Interpretation Agent* is the most likely intention expected by the computational agent and expressed by the human-user, on the basis of the actual image or linguistic message, and the contents stored in memory.

2.1 Dialogue Manager Agent

The *Dialogue Manager agent* is the central component of this architecture. It is responsible of managing the conversation between the human user and the robot. The dialogue manager is a program interpreter that interprets objects called *dialogue models*. These specify conversational protocols in terms of conversational situations, which in turn are specified in terms of a set of the meaningful intentions that can plausible be expressed by the human-user in the situation (i.e. the expected intentions) with their associated actions. Dialogue models are represented as functional recursive transition networks [2], where nodes symbolize conversational or interactive situations (for example: *listening*, *telling* or *recursive*), and arcs are labeled by the intentions that need to be expressed to reach the corresponding situation (called *input speech acts*), and by the actions that are performed when such situations are reached (called *output rhetorical acts*), see Fig. 1(c). Arcs can be labeled with constants, grounded predicates or functions that have the current situation and the conversational history as their arguments,

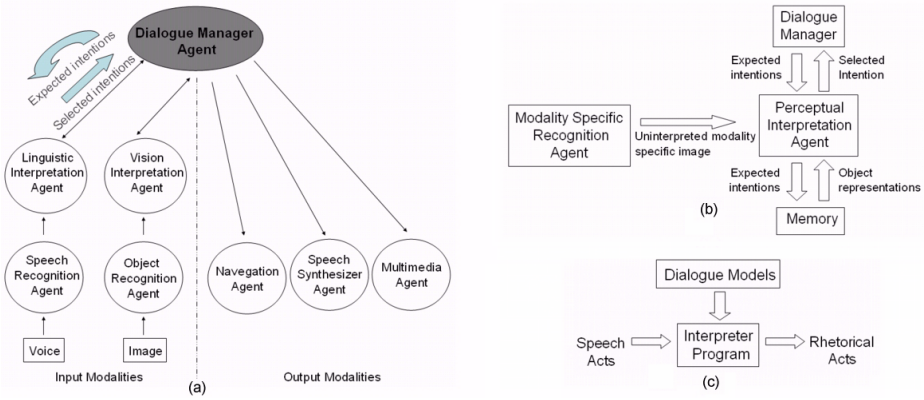


Fig. 1. (a) Agent's architecture, (b) General structure of an input modality module, (c) Dialogue Manager

and are evaluated on the fly when such arcs are traversed. Dialogue models are modular self contained objects, and are handled by the dialogue manger through a stack discipline. The interpretation of dialogue models process runs hand in hand with the interaction of the robot with the world.

2.2 Language Perception Module

The language perception module, in turn, assigns an interpretation to the text produced by the *Speech Recognition Agent*. For this, every expected intention in every interpretation situation has an associated regular expression, which states an ample collection of forms (complete and partial) through which such intention can be expressed by the human user. If the text matches the regular expression, the corresponding intention is selected, and its associated action is performed. In the case that none of the expected intentions can be matched, the system loads and executes a recovery dialogue model which may involve linguistic, visual and even motor behavior. The speech recognizer is an in house system built with *Sphinx*[1] and the Corpus DIMEX100 for Mexican Spanish, [6].

2.3 Vision Perception Module

Vision tasks include face recognition, gesture recognition, vision-based localization, and object recognition, for example. In this paper we focus in this latter task. There are four traditional approximations to object recognition (even though most of the systems use several of these techniques): 1) appearance based methods, 2) grammatical correspondence and graphs, 3) geometric based methods, and 4) local invariant feature correspondence. During the last years the approximation based on invariant local features has been widely accepted, [7][8].

[1] http://www.speech.cs.cmu.edu/

Here we explore the use of the combination of SIFT local feature descriptors [9], the Best Bin First (BBF) matching algorithm [10] and the Graph Transformation Matching (GTM) outlier removal algorithm [11], for the interactive object recognition problem on a mobile robot.

First, SIFT interest feature points are detected within an image. Each interest point detected is defined by its pixel position in the 2D image, a scale, an orientation, and a 128 feature vector describing the keypoint's surroundings. Known objects are defined in a database composed of object images and their corresponding keypoint descriptors and labels. The object recognition process consists in finding a correct match between keypoint descriptors of the actual robot view image and the known objects stored in the database. For this, a combination of the BBF and GTM algorithms was used. BBF is an algorithm that efficiently finds an approximate solution to the nearest neighbor search problem. A disadvantage of this algorithm, is that there is no spatial information taken into consideration (other than region similarity) for obtaining the nearest match. This causes the introduction of erroneous matches called outliers. The GTM algorithm was introduced here to reduce these outliers. GTM has a simple but effective conducting principle: iteratively eliminate correspondences

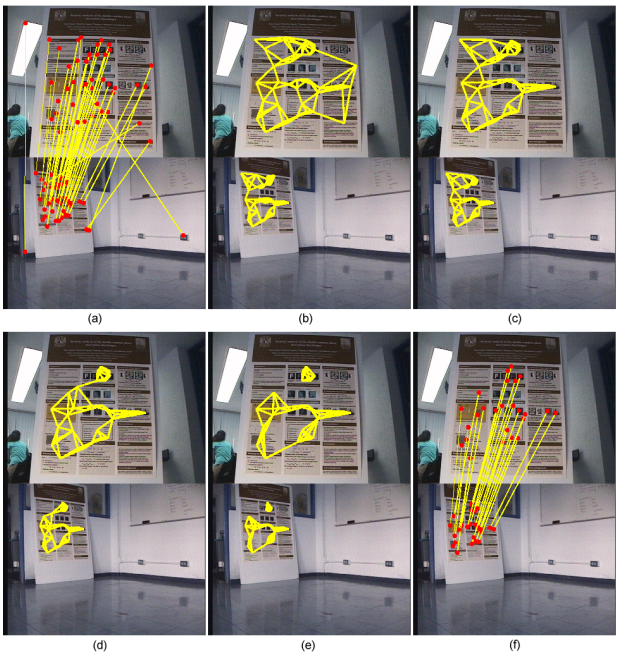


Fig. 2. Example of feature matching results for object recognition: a) BBF initial matching, b) GTM graphs at iteration 1, c) graphs at iteration 6, d) graphs at iteration 12, e) GTM final identical graphs at iteration 17, f) GTM final matching

that disrupt the neighborhood relationships. In order to do so, it constructs a K-nearest-neighbor (K-NN) graph for each view image (based on the space coordinates of detected feature points), and during each iteration it: a) removes the vertex (match) which most disrupt the similarity structure on both graphs, b) reconstructs the corresponding K-NN graphs and repeats this process until both graphs are identical, which means outliers have been removed [11]. Fig. 2.3(a) shows an example of the matching result from the BFF algorithm. Fig. 2.3(b)-(e) shows the graph transformation process through iterations 1, 6, 12 and 17. The final GTM matching (after outliers removal) is shown in Fig. 2.3(f).

As in the case of the language perception module, this module is responsible of trying to figure out what is the most probable object it is seeing, given the known objects knowledge, the actual robot view evidence and the expected intentions in the current dialogue model. This module is again composed of two agents: 1) *Object Recognition Agent*, and 2) *Vision Interpretation Agent*. The *Object Recognition Agent* takes the actual robot view, and describes it in terms of it's SIFT descriptors. The *Vision Interpretation Agent* uses this SIFT descriptors and tries to match them (using the BBF and GTM algorithms) just with the known objects in the list of expected intentions, which highly reduce the search space. The output of the *vision perception module* is the label of the recognized object. This label is passed back to the dialogue manager in order to start a conversation related to what the robot is seeing. If this module could not recognize any object, the dialogue manager starts a recovery conversation protocol.

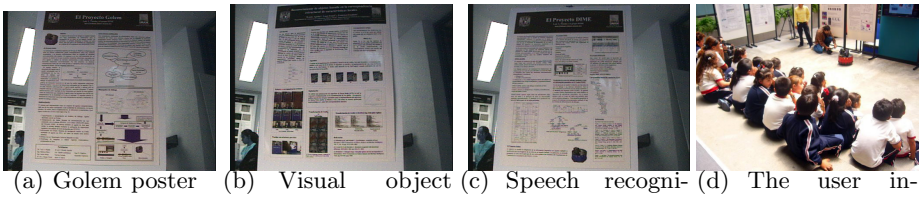
2.4 Output Modality Agents

In addition to the input modality agents, the system supports a number of output modalities, which are associated to the actions performed in response to the interpretation of the human-user intentions. Actions are defined in multimodal rhetorical structures, which are defined in the dialogue models. Rhetorical structures (e.g. introduction, presentations, explanation, elaboration, etc.) are composed of basic rhetorical acts, that are modality specific actions, like pronouncing an utterance or displaying a text or an image; even the robot's motor behavior is specified as a basic rhetorical act. When a rhetorical act is performed all its basic acts are dispatched sequentially but performed simultaneously. Spoken acts are realized through a speech synthesizer. This agent is implemented using *Festival*[2]. There is also a multimedia agent that is responsible of displaying complementary visual information like text, images, animations or videos to the user. This agent is implemented in *Java* using the *Java Media Framework*[3]. This architecture was implemented using the Open Agent Architecture (OAA)[4] on a Magellan Pro robot - named *Golem*.

[2] http://www.cstr.ed.ac.uk/projects/festival
[3] http://java.sun.com/javase/technologies/desktop/media/jmf/index.jsp
[4] http://www.ai.sri.com/oaa/

3 Tests and Results

In the present application Golem performs as the guide of a posters session about the research projects in our computer science department. When a visitor arrives, the main dialogue is executed, and the robot starts the conversation by introducing himself and guiding the user through the possible conversational topics. There are two ways of selecting the topic of conversation: 1) ask the robot for a specific project through speech, or 2) show to the robot the poster of the project you are interested in. If the user chooses to interact through vision, then it has to take the selected poster and put it on a position where the robot can see it. For the second approach, an object database needs to be created. A frontal picture for each poster was taken with resolution of 320×240 pixels. No considerations were taken about illumination or position. The complete database is composed of 13 poster images, and its corresponding SIFT descriptors and labels. A subset of 3 poster images is shown in Fig. 3(a)-(c).

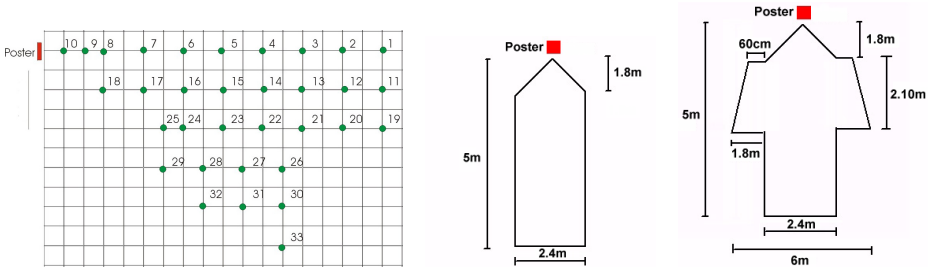


(a) Golem poster (b) Visual object recognition poster (c) Speech recognition project poster (d) The user interacting with the robot

Fig. 3. (a)-(c) A subset of the object image database, (d) Presentation of the robot Golem in a museum

The following is the English translation (from the original in Spanish) of a typical dialogue of the second kind:

GOLEM: *Please, show me the poster that you would like to talk about, and give me a second to recognize it.*
GOLEM: *Hey, it's my poster! Excellent choice.*
GOLEM: *About my poster I can tell you who are my programmers, how I work, or how am I physically. Are you interested in something like this, or do you prefer another poster?*
USER: *Tell me about how you are physically, please.*
GOLEM: *Physically I'am composed of 16 sonars, ... I can also talk to you about who are my programmers or how I work. Are you interested on these topics or do you prefer to talk about another poster?*
USER: *I'd better like to talk about another poster.*
GOLEM: *Ok. Show me the poster you are interested in so I can see it.*
GOLEM: *I cannot recognize this poster. I'm showing you on the screen what my actual view is. Please, move it to a better position.*
GOLEM: *Hey, I can see it now! It's the speech recognition poster ...*



(a) Experiment setting map. The size of each square is 30 × 30cm

(b) Safety zone with just one poster frontal view in the database

(c) Safely zone with 3 poster views: one frontal and two lateral views at 35 degrees

Fig. 4. Object recognition test map and results

A series of systematic experiments were performed to test the robustness of the object recognition module. Fig. 4(a) shows part of the laboratory where the guided visit takes place. All the 13 posters were set at the position marked with a rectangle (one at a time), and the robot was asked to recognize the poster on each of the 33 circles. This was performed with three different illumination conditions: 1) in the morning, 2) in the afternoon and 3) in the night with artificial light. The experiment was repeated changing the orientation of the poster from 0 to 360 degrees of rotation on the same plane. Figure 4(b) shows the results of this experiment when using just one frontal view of the posters in the database. The figure shows the zone where the robot recognized correctly the poster always. This test showed that the farthest recognition distance is 5m and the closest is 60cm. When two more views of the poster were added to the database (taken at 30cm away from the frontal view, one to the left and one to the right), the safety zone grew as shown in Fig. 4(c). About the execution times, the interaction between the user request for the explanation of a specific poster and the robot recognition answer is about 5 seconds.

This application has been demonstrated on more than 50 guided visits to the department, with children and adult users. It has also been taken outside the lab to be tested on talks about this project, where the illumination and sound conditions are natural and uncontrolled. Fig. 3(d) shows a picture of the presentation of this robot at the Museum of Science of Mexico City.

4 Conclusions and Future Work

This paper presented a successful integration of a graph-based perception module into a three level agent architecture centered on the dialogue. The architecture was implemented and tested on a service robot which acts as a visitor's guide of our department. Vision was used for interactive object recognition, where a graph transformation matching algorithm was used for feature matching.

Experiments and results of the object recognition module were presented and a safely recognition zone was defined. This application has been demonstrated on more than 50 guided visits to our department, with children and adult users. As future work we are incorporating new visual tasks, such as gesture recognition and visual-based localization.

References

1. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A.: An architecture for a generic dialogue shell. Natural Language Engineering 6(34), 213–228 (2000)
2. Pineda, L.A.: Specification and interpretation of multimodal dialogue models. In: Sidorov, G. (ed.) Memorias del Workshop de robots de servicio, MICAI (2008)
3. Wachsmuth, S., Fink, G.A., Kummert, F., Sagerer, G.: Using speech in visual object recognition. In: Mustererkennung 2000, 22. DAGM-Symposium Kiel, Informatik Aktuell, pp. 428–435. Springer, Heidelberg (2000)
4. Saenko, K., Darrell, T.: Towards adaptive object recognition for situated human-computer interaction. In: Proceedings of the 2007 Workshop on Multimodal Interfaces in Semantic Interaction, pp. 43–46 (2007)
5. Rahmadi, K., Altab, H.M., Akio, N., Yoshinori, K.: Object recognition through human-robot interaction by speech. In: 13th IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN, pp. 619–624 (2004)
6. Pineda, L.A., Villasenor, L., Cuétara, J., Castellanos, H., López, I.: Dimex100: A new phonetic and speech corpus for mexican spanish. In: Lemaître, C., Reyes, C.A., González, J.A. (eds.) IBERAMIA 2004. LNCS, vol. 3315, pp. 974–983. Springer, Heidelberg (2004)
7. Obdrzálek, J.M.: Object recognition methods based on transformation covariant features. In: XII European Signal Processing Conference EUSIPCO 2004, pp. 1333–1336 (2004)
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
9. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. Beis, J., Lowe, D.: Shape indexing using approximate nearest-neighbour search in highdimensional spaces. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, pp. 1000–1006 (1997)
11. Aguilar, W., Frauel, Y., Escolano, F., Pérez, M.M., Espinosa-Romero, A., Lozano, M.: A robust graph transformation matching for non-rigid registration. Image and Vision Computing (2008), doi:10.1016/j.imavis.2008.05.004