INTEGRATING POINTING GESTURES INTO A SPANISH–SPOKEN DIALOG SYSTEM FOR CONVERSATIONAL SERVICE ROBOTS

Héctor Avilés, Iván Meza, Wendy Aguilar, Luis Pineda

Department of Computer Science,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México,
Circuito Escolar, Ciudad Universitaria, D.F.
04510 México
haviles@turing.iimas.unam.mx, imeza@turing.iimas.unam.mx, weam@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

Keywords: Conversational service robots, dialog systems, pointing gesture recognition, multimodal interaction.

Abstract: In this paper we present our work on the integration of human pointing gestures into a spoken dialog system in Spanish for conversational service robots. The dialog system is composed by a dialog manager, an interpreter that guides the spoken dialog and robot actions, in terms of user intentions and relevant environment *stimuli* associated to the current conversational situation. We demonstrate our approach by developing a tour–guide robot that is able to move around its environment, visually recognize informational posters, and explain sections of the poster selected by the user *via* pointing gestures. This robot also incorporates simple methods to qualify confidence in its visual outcomes, to inform about its internal state, and to start error–prevention dialogs whenever necessary. Our results show the reliability of the overall approach to model complex multimodal human–robot interactions.

1 INTRODUCTION

We present the integration of pointing gestures into a Spanish–spoken dialog system for conversational service robots. The main component of the dialog system is the dialog manager that interprets task–oriented dialog models which define the flow of the conversation and the robot actions. The dialog manager, an agent itself, also coordinates other distributed agents that perform speech, navigation and visual capabilities for the robot in terms of user intentions and perceptual *stimuli* relevant to the current conversational situation (Aguilar and Pineda, 2009).

To demonstrate our approach, we developed a tour–guide robot that asks for one of six informational posters using spoken language, navigates to the poster to recognize it and identifies its sections. The robot is able to explain sections selected by the user *via* 2D pointing gestures. We also explore error prevention and recovery dialogs for the visual system, by incorporating a simple method to qualify confidence of the robot in its visual outcomes, to begin confirmation dialogs whenever necessary.

2 RELATED WORK

Robita (Tojo et al., 2000) is a robot that is able to recognize questions about the location of a person in an office environment, to answer verbally sentences, to point places with its arm, and also recognize pointing gestures of the user. ALBERT (Rogalla et al., 2002) is a robot capable to grasp objects following speech and pointing gestures. Jido (Burger et al., 2008) tracks head and hands of a user, and answers requests such as “Come here”. Markovito (Aviles et al., 2009) recognizes spoken commands and identifies 9 gestures of the type “come” or “attention”. In these examples, dialog modules are subordinated to the requirements of a main coordinator module. In this form, they cannot be considered conversational service robots. Only few robots present dialog systems as the core element of their coordination modules. BIRON (Toptsis et al., 2005) is a robot that uses pointing gestures and visual information of the environment to resolve spoken object references. The dialog module is a finite state machines. Similarly, the Karlsruhe robot (Stiefelhagen et al., 2006) includes a dialog manager based on reinforcement learning.

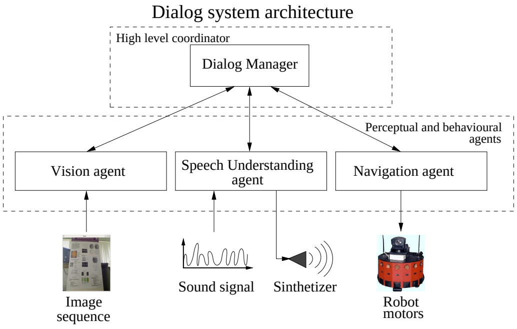


Figure 1: Architecture of the dialog system.

3 DIALOG SYSTEM

The architecture of our dialog system is depicted in Figure 1. This is a 2–layer architecture composed by four independent agents: i) the dialog manager, ii) the Vision agent, iii) the Speech Understanding agent, and iv) the Navigation agent. The dialog manager, as the high level coordinator, executes instructions pre–defined in a dialog model designed for the task at hand.

3.1 Dialog manager

The dialog manager is an interpreter based on *recursive networks* that tracks the context of the conversation, manages the set of perceptual information for a given stage of the interaction, and also produces adequate responses for the current conversational situation. It is also responsible of mediating the agent level communication between perceptual and behavioral agents. The dialog manager performs all these functions by executing task–oriented dialog models.

Dialog models specify interaction protocols of the robot with the user and environment. The core elements of a dialog model are: i) *situations* that represent a relevant state in interaction for which a particular perceptual strategy has to be applied –*e.g.*, listening to the answer of a question posted by the robot. ii) *expected intentions*, or the set of perceptual stimuli relevant for a given situation iii) *multimodal rhetorical structures* or *MRS*, that represent a set of basic rhetorical acts –or behaviours– to be performed by the system –*e.g.*, speech synthesis or robot motion. These three elements allow us to codify a rich set of behavioural capabilities and perceptual stimuli that our robot is able to understand.

Figure 2 graphically describes one fragment of the dialog model developed for our robot. The nodes represent the situations, the labels on the arcs represent the pair of expected intention and MRS structure, and the arrows point to the destination situation. The dialog model starts with the situation (n_1) in which the system presents the posters. In this case, the rhetorical act to perform depends on the history of the dialogue. We achieve this with the evaluation function represented by: $get(H,D,Posters)$ where the returned value of this function is a MRS structure enumerating the poster which have not been visited. The next situation is a listening situation (l) in which the poster to visit is interpreted from the user speech. At this situtation, there are two options: the poster is valid or invalid. When an invalid poster is chosen the robot will mention this and return to the situation n_1. In case the chosen option is valid the robot moves to the poster location ($goto(Poster)$). In this case, it reaches the situation seeing (s) where it will look at the poster and identify it. The next situation will be defined by a functional evaluation of the seen poster being the right one or not $validate(H,D,Poster)$. Depending on the result the dialogue will reach the situation n_2 in which it will explain the poster or the recursive situation r in which it goes into a sub–dialog to figure out what it went wrong with the poster.

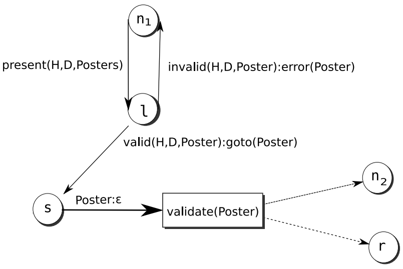

n1
present(H,D,Posters)
invalid(H,D,Poster):error(Poster)
l
valid(H,D,Poster):goto(Poster)
s
Poster:ε
validate(Poster)
n2
r

Figure 2: Example of one segment of our dialog model.

4 PERCEPTUAL AND BEHAVIOURAL AGENTS

4.1 Speech Understanding Agent

The Speech Understanding agent performs speech recognition and synthesis in Spanish (Pineda et al., 2004), and speech interpretation, that is performed by comparing the sentence with a set of linguistic patterns of each expected intention.

4.2 Navigation Agent

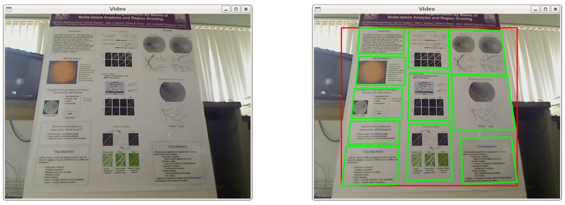
Robot navigation is performed on a 2D world. The dialog manager informs to the agent the world (x,y)

position of the poster plus θ, the final angular orientation. First, the robot rotates accordingly to the actual orientation and final (x,y) location, and moves along a straight line up to this coordinate. We assume there are not obstacles along the path of the robot. Once the robot has arrived, it rotates again up to reach the final θ orientation.

4.3 Vision Agent

Poster recognition is performed using SIFT algorithm. For recognition, features are obtained from the current view and matched against each SIFT poster template. The number of matches of each poster is stored in a frequency table R which is used to qualify visual outcomes as described in section 5.1. The poster with a maximum number of matches and above a threshold –defined experimentally– correspond to the classification result. If this criterion is not met, the visual agent let the dialog manager know this situation, and a simple recovery dialog proceeds.

To carry out region segmentation, the vision agent receives (i,j) original coordinates of the rectangle that delimits each section of the poster. To adjust coordinates to the actual view, we calculate a perspective transformation matrix based on SIFT matches using RANSAC. Once all visible sections have been calculated, a rectangular window of interest is defined to fit all of them. Figure 3a and 3b shows the original view of the poster, and its relevant regions and poster window, respectively. From now on, visual analysis will be confined to this window and referred simply as the poster.



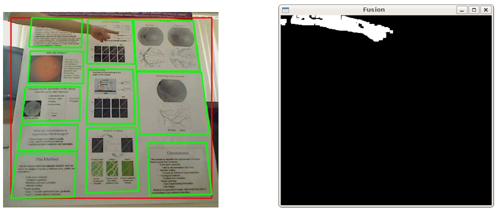
(a) (b)

Figure 3: Examples of region segmentation: a) original image, b) segmentation of each relevant region of the poster. Region boundaries are delimited by green rectangles. The red rectangle defines the poster window.

For pointing gestures, the arm is spotted into a binary image F by fusioning motion data –image M– the difference between edges using Laplacian edge detectors –image E– of the current poster view and the first image of the poster taken in the actual robot's position, and the absolute difference of the current poster image and its initial view –image D. These 3 images are thresholded to get binary images. Data fusion is performed following the logic *AND* rule:

$$F(i,j) = M(i,j) \wedge E(i,j) \wedge D(i,j), \ \forall i,j, \quad (1)$$

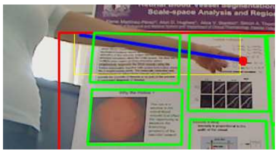
Figure 4 the original monocular image and the resulting fusion image F.



(a) (b)

Figure 4: Results of the fusion of simple visual clues for arm detection: a) original image with the poster window drawn in black color, and b) fusion results.

Arm segmentation is executed by scanning F row–by–row, from left–to–right, and from top–to–bottom, to detect foreground–background pixels. Simple decisions are used to grow and identify the foreground region of the arm. A line is fitted to all its pixels using least–squares method. The tip of the arm is selected by comparing the distance from both extreme points of the line to the vertical edges of the poster. Figure 5 shows the visual outcome of this procedure. One vector P is used to record the number of image frames that the tip is over each region. Each bin of P correspond to a single region. The first region to accumulate 30 images, is considered to be the user choice. Arm segmentation is performed for a maximum of 10 seconds. In case a section is not identified by the end of this period of time, the dialog manager is informed so it could start a sub–dialog with the user.



(a)

Figure 5: Result of the arm segmentation, line fitting and tip selection.

5 PUTTING IT ALL TOGETHER

5.1 Evaluation of Visual Outcomes

To evaluate poster recognition results, we use table *R* described above. We averaged and plotted tables *R* obtained from more than 300 classification trials. It was observed regular patterns for correct and incorrect classification outcomes. We propose to model these patterns using Shannon's entropy measure *H*. When a poster is recognized, *R* is normalized to obtain a discrete probability distribution, and *H* is calculated for this distribution. We consider three main categories to evaluate the confidency on poster classification: *high*, *medium* and *low* assigned accordingly to threshold values of *H* defined by experimentation.

To evaluate the selection of a region we use vector *P* described in Section 4.3. If the bin with the maximum number of image frames is above 15 and below 30, then the robot warns the user that it is not completely sure about the selection. If the bin with the maximum is below 15, the robot tells the user that the identification of a section was not possible, and asks for another attempt.

5.2 Results

We have tested our approach with 5 different people in several demonstrations in our Lab. All these people are either students or Professors of our Department. In all cases, the robot was able to correctly identify the desired poster, or to ask the user in case of doubt. Almost all people were able to select the desired section of the poster within a single trial, and they seemed to be satisfied with the corresponding explanations. However, we also observed that not all people points to the poster immediately after the alert sound is emitted by the robot, mainly because they had not yet decided which section to select. For this case, evaluation of the pointing output has proved to be an useful tool to add flexibility to our system. From initial usability tests performed with these users, we found that evaluating confidence of the visual analysis improves considerably the perceived naturalness of the spoken language of the robot.

6 CONCLUSIONS

In this paper we presented our work on the integration of pointing gestures into a spoken dialog system in Spanish for a conversational service robot. The dialog system is composed by a dialog manager, that interprets a dialog model which defines the spoken dialog and robot actions, accordingly to the user intentions and its environment. We developed a tour–guide robot that navigates in its environment, visually identify informational posters, and explain sections of the poster pointed by the user with its arm. The robot is able to qualify its confidence in its visual outcomes and to start error–prevention dialogs with the user. Our results showed the effectiveness of the overall approach and the suitability of our dialog system to model complex multimodal human–robot interactions.

REFERENCES

Aguilar, W. and Pineda, L. (2009). Integrationg Graph–Based Vision Perception to Spoken Conversation in Human–Robot interaction. In *10th International Work–Conference on Artificial Neural Networks*, pages 789–796.

Aviles, H., Sucar, E., Vargas, B., Sanchez, J., and Corona, E. (2009). *Markovito: A Flexible and General Service Robot*, chapter 19, pages 401–423. Studies in Computational Intelligence. Springer Berlin / Heidelberg.

Burger, B., Lerasle, F., Ferrane, I., and Clodic, A. (2008). Mutual assistance between speech and vision for human–robot interaction. In *IEEE/RSJ International Conference on Intelligent Robotics and Systems*, pages 4011–4016.

Pineda, A., Villaseñor, L., Cuetara, J., Castellanos, H., and Lopez, I. (2004). Dimex100: A new phonetic and speech corpus for Mexican Spanish. In *Iberamia-2004, Lectures Notes in Artificial Intelligence 3315*, pages 974–983.

Rogalla, O., Ehrenmann, O., Zoellner, M., Becher, R., and Dillmann, R. (2002). Using gesture and speech control for commanding a robot assistant. In *Proceedings of the 11th IEEE Workshop on Robot and Human interactive Communication*, pages 454–459.

Stiefelhagen, R., Ekenel, H., C. Fugen, P. G., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., and Waibel, A. (2006). Enabling multimodal human–robot interaction for the karlsruhe humanoid robot. In *Proceedings of the IEEE Transactions on Robotics: Special Issue on Human–Robot Interaction*, pages 1–11.

Tojo, T., Matsusaka, Y., and Ishii, T. (2000). A Conversational Robot Utilizing Facial and Body Expressions. In *International Conference on Systems, Man and Cybernetics, (SMC2000)*, pages 858–863.

Toptsis, I., Haasch, A., Hüwel, S., Fritsch, J., and Fink, G. (2005). Modality integration and dialog management for a robotic assistant. In *European Conference on Speech Communication and Technology*, Lisboa, Portugal.