SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 190–191.
Los Angeles, California, 21–23 September 2011.

A Smart Interaction Device for Multi-Modal Human-Robot Dialogue

Glenn Taylor, Richard Frederiksen, Jacob Crossman, Jonathan Voigt, and **Kyle Aron**
SoarTech
3600 Green Court Suite 600
Ann Arbor, MI 48105
{glenn,rdf,jcrossman,jon.voigt,aron}@soartech.com

Abstract

This paper introduces a Smart Interaction Device (SID) that enables a multi-modal dialogue between a user and a robot to help reduce the operator's workload in performing complex robot tasks. We describe SID and a demonstration of its performance in a robot navigation task.

1 Smart Interaction Device

Most user interfaces for ground robots are Operator Control Units (OCUs) that require significant heads-down time to operate and involve giving the robot detailed low-level tasks. We present a Smart Interaction Device (SID) whose purpose is to make the user's interaction more natural, requiring less work. Specifically, SID enables users to interact with a robot using speech and pointing gestures to accomplish tasks.

Our approach is to introduce a smart interface layer between the user and the robotic system. As shown in Figure 1, SID consists of a reusable core ("SID Core") that manages a dialogue with the user, translates user intent into robot terms, and can monitor robot progress against the user's intent. Additionally, SID uses plug-ins for input-specific, and platform-specific layers, each of which may be customized to a particular application. Different ways of interacting with the robot and different robot APIs necessitate different user-facing and robot-facing software interfaces. We have connected SID to two different robotic platforms (air, ground) and a UAV simulation environment, and have connected to an iPhone and Microsoft Kinect for gesture inputs.

SID Core is implemented using the Soar cognitive architecture (Laird, Newell, & Rosenbloom, 1991), which gives us a robust platform for knowledge-based reasoning. In this system, Soar is used for reasoning about dialogues and tasks, where different kinds of knowledge are put to different uses. Soar provides a framework for uniform representation of knowledge (rules) and fast application of that knowledge using a Rete matching algorithm. We also take advantage of some newer features of Soar, such as query-accessible Semantic Memory.

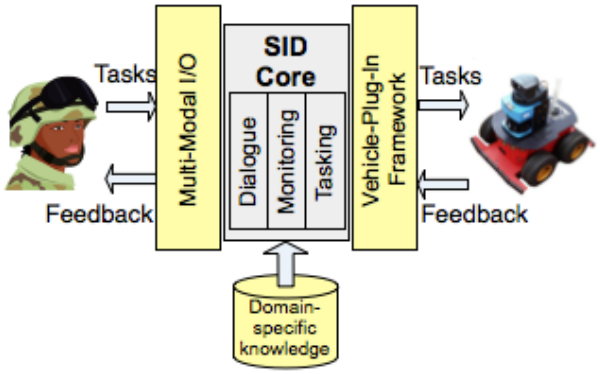


Figure 1: High Level Architecture of the Smart Interaction Device (SID)

Individual input modes are recognized independently, and converted into semantic frame representations. Multiple input modes are combined via frame-based unification. When the user provides new input, the semantic frame generated by the speech, gesture, or their combination, is stored as a dialogue move. The dialogue move is classified based on the taxonomy of (Traum, 2003), using domain-specific rules that look at the content of the input and the current dialogue context. With this classification, SID's DM then determines whether this dialogue move is part of an existing dialogue (does it share the same topic?), or whether it is the start of a new dialogue (is it a new command?). Once the dialogue move is

assigned to a dialogue, the system can begin resolving references within the user's input.

Resolving references to objects in the environment is a search problem looking for objects with features described by the user. If there is a single unique match, then the system can simply use the object's location as the destination. If there is no such object retrieved or if there are multiple objects retrieved, then the system must ask for clarification. This request from the system is a dialogue move that starts a sub-dialogue to request clarification from the user. With a complete command, SID can then generate tasking for the robot. In general design, SID resembles the WITAS System 1 (Lemon, Bracy, Gruenstein, & Peters, 2001), but with the addition of 3D pointing gestures.

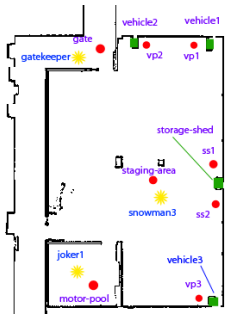
2 Prototype

From these concepts, we have developed an end-to-end prototype that lets human users and a robot interact in mobility tasks. We use a MobileRobots P3AT Pioneer robot with forward-looking LIDAR as the primary sensor. The robot has a pre-built map of the task area with hand-annotated objects and location names. This map is used for resolving references from the user and navigation. The objects primarily consist of cardboard boxes as stand-ins for "vehicles" or "buildings" that could be referred to. The on-board robot capabilities allow for planning routes to x-y locations, avoiding obstacles as needed. It can also be given low-level movement commands such as move forward/backward, turn left/right, and stop.

With the addition of SID, the robot's capabilities are extended to taking inputs via speech and pointing gestures. The current system is speech-dominant: gestures serve primarily to disambiguate or clarify verbal utterances. In cases where a gesture is not given or the gesture recognizer fails to register a gesture, the system can ask for clarification. For example:

User: *"Go to that vehicle" (no gesture)*
System: *"Which vehicle? I know of a blue vehicle and a red vehicle."*
User: *"That one." (pointing to the blue vehicle)*
System: *"Okay, going to the blue vehicle."*

We use an iPhone as the primary input and output device for the user, which serves as a simulated radio (speech input and output) and a pointing device (gesture input). Speech recognition is performed off-board the iPhone using a COTS recognizer with grammar-based recognition, the output of which is then passed through a semantic parser. Gesture recognition and speech generation both occur on the device itself. Both speech and gesture are enabled via a push-to-communicate button to reduce the amount of errant input.


(a)


(b)

Figure 2: (a) A user gesturing while speaking to the robot; (b) the map of the task area (10m x13m) with labeled locations and objects

In addition to tasking the robot to move, the user can request status such as robot location and current task, and can request to be informed when the robot completes a task. With these kinds of information requests, the user does not have to constantly attend to the robot while it is performing a task. This is one key feature that helps separate SID from the standard OCUs: rather than staring at OCUs to task a robot, users of SID-enabled robots can perform other tasks and maintain awareness of their surroundings while tasking the robot.

Acknowledgments

This work was partially funded under ONR Contract #N00014-10-M-0403.

References

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1991). Soar: An Architecture for General Intelligence. *Artificial Intelligence, 47*, 289-325.

Lemon, O., Bracy, A., Gruenstein, A., & Peters, S. (2001). *The WITAS multi-modal dialogue system 1*. Paper presented at the Proc. European Conference on Speech Communication and Technology.

Traum, D. (2003). *Semantics and Pragmatics of Questions and Answers for Dialogue Agents.* Paper presented at the International Workshop on Computational Semantics.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 192–193.
Los Angeles, California, 21–23 September 2011.

Effects of 2D and 3D Displays on Turn-taking Behavior in Multiparty Human-Computer Dialog

Samer Al Moubayed **Gabriel Skantze**
KTH Speech Music and Hearing
Stockholm, Sweden
sameram@kth.se,gabriel@speech.kth.se

Abstract

The perception of gaze from an animated agent on a 2D display has been shown to suffer from the Mona Lisa effect, which means that exclusive mutual gaze cannot be established if there is more than one observer. In this study, we investigate this effect when it comes to turn-taking control in a multi-party human-computer dialog setting, where a 2D display is compared to a 3D projection. The results show that the 2D setting results in longer response times and lower turn-taking accuracy.

1 Introduction

The function of gaze for interaction purposes has been investigated in several studies. Gaze direction and dynamics have been found to serve several different functions, including turn-taking control, deictic reference, and attitudes (Kendon, 1967). Recently, there has been an increasing interest in virtual agents that may engage in multi-party, situated dialogue (e.g., Bohus & Horvitz, 2010). In such settings, gaze may be an essential means to address a person in a crowd, or pointing to a specific object out of many.

It is known that perception of 3D objects that are displayed on 2D surfaces is guided by, what is commonly referred to as, the Mona Lisa effect (Todorovic, 2006). This means that the orientation of the 3D object in relation to the observer will be perceived as constant, no matter where the observer is standing in the room. This effect has important implications for the design of interactive systems, such as embodied conversation agents, that are able to engage in situated interaction, as in pointing to objects in the environment of the interaction partner, or looking at one exclusive observer in a crowd.

In a previous study (Al Moubayed et al., in press), we have measured how subjects *perceive* gaze direction using an animated agent in 2D and 3D conditions (see Figure 1). The purpose of this study is to investigate how gaze may affect the turn-taking *behavior* of the subjects in a multi-party human-computer dialog, depending on the use of 2D or 3D displays.

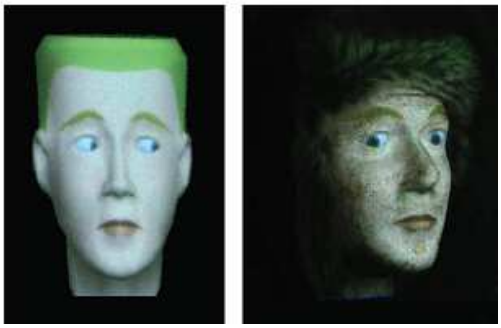


Figure 1: A snapshot of the animated agent projected on a 2D white board (left) and on a 3D head model (right).

2 Method

Two sets of five subjects were asked to take part in the experiment. In each session, the five subjects were seated at fixed positions at an equal distance from each other and from an animated agent. The agent addressed the subjects by directing its gaze in their direction. Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface (using the same 3D computer model), as shown in Figure 1. The

conversational behavior of the animated agent was controlled using a Wizard-of-Oz setup. For each new question posed by the agent, the gaze was randomly shifted to a new subject. The subjects were given the task of watching a first-person video from a camera navigating around the city of Stockholm, after which the animated agent asked them to describe the route they had just seen. After each video was finished, the animated agent started to ask the subjects about directions on how to reach the landmark the video ended with, starting from the point of view the video started with. The dialogs were in Swedish. An example dialog is shown in Table 1. Each set of subjects did four dialogs in each 2D/3D condition (i.e., a total of eight videos). The ten subjects were not informed about the purpose of the experiment beforehand.

Table 1: The beginning of one of the recorded interactions translated from Swedish.

Agent	Hi, can you help me to find my way to the Finnish church?
Subject 1	Yes
Subject 2	Yes
Agent	Should I continue in this direction?
Subject 3	Yes, a little bit at least
Agent	How far should I go then?
Subject 4	I think it was about 20 meters, maybe a bit longer
Agent	Is there some landmark I should look for?
Subject 2	You should look for the Great Church, when you come to Västerlånggatan you will see the Great Church on your left, then should pass it
Agent	Do you agree?
Subject 5	Yes

3 Results

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. The mean response time was also calculated for each condition, i.e. the time between the gaze shift of the question and the time takes for one of the subjects to answer. A two sample ANOVA analysis was applied, with the response time as a dependent variable, and the condition as an independent variable. The results show a significant main effect [$F(1)=15.821$, $p<0.001$], with a mean response-time of 1.85 seconds for the 2D condition, and of 1.38 seconds for the 3D condition. No significant correlation with time was found (Pearson Correlation = -0.094), which means that there is no learning effect on how to perceive the gaze of the agent for either condition.

4 Conclusions

The results show that the use of gaze for turn-taking control on 2D displays is limited due to the Mona Lisa effect. The accuracy of 50% is probably too low in settings where many users are involved. By using a 3D projection, this problem can be avoided to a large extent. However, the accuracy for the 2D condition was higher than what was reported in a previous perception experiment in a similar setting (Al Moubayed et al., in press). A likely explanation for this is that the subjects in this task may to some extent compensate for the Mona Lisa effect – even if they don't "feel" like the agent is looking at them, they may learn to associate the agent's gaze with the intended target subject. This comes at a cost, however, which is indicated by the longer mean response time. The longer response time might be due to the greater cognitive effort required making this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

Acknowledgements

This work has been carried out at the Centre for Speech Technology at KTH, and is supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314, as well as the SAVIR project (Situated Audio-Visual Interaction with Robots) funded by the Swedish Government (strategic research areas).

5 References

Bohus, D. & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of ICMI-MLMI*, Beijing, China.

Al Moubayed, S., Edlund, J., & Beskow, J. (in press). Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems.*

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Todorovic, D. (2006). Geometrical basis of perception of gaze direction. *Vision Research, 45*(21), 3549-3562.