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 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.

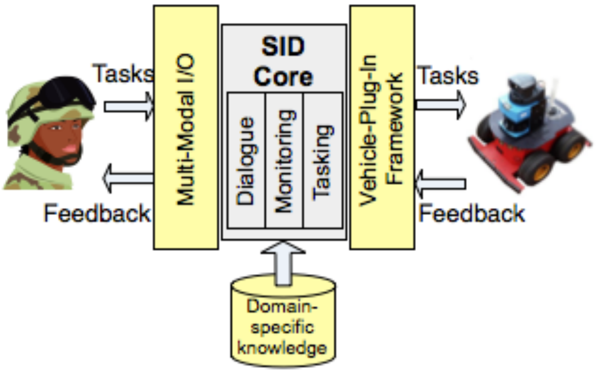

Figure 1: High Level Architecture of the Smart Interaction Device (SID)

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.

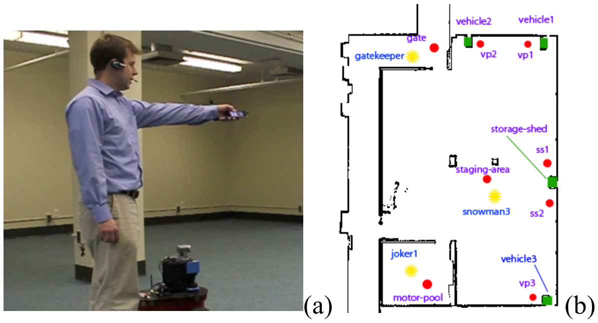

Figure 2: (a) A user gesturing while speaking to the robot; (b) the map of the task area (10m x13m) with labeled locations and objects

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.