

Reference Resolution in Multimodal User Interfaces to Map-Based Applications

Lisa Harper • Dan Loehr

703-983-5241 703-983-6765
lisah@mitre.org loehr@mitre.org

MITRE Sponsored Research

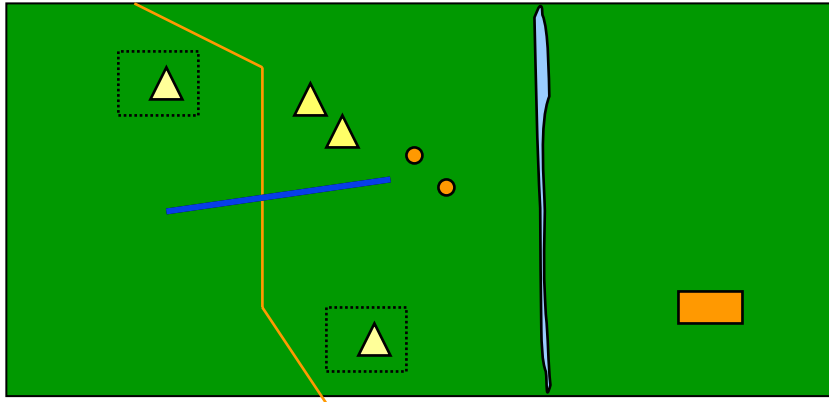
The logo for the MITRE Technology Program, featuring a stylized graphic of stacked blocks in yellow, orange, and blue to the left of the text.

MITRE
Technology
Program

Problem

- The world is moving towards "ubiquitous computing," or the notion of using small portable computing devices everywhere.
- While the hardware for these devices is maturing rapidly, the user interfaces are not. They will require a multimodal synthesis of text, gesture, and speech beyond the current state of the art.
- One of the problems faced by such interfaces is the need to resolve *references* (what the user is referring to) across modalities. A user may refer to an item in a display by using speech ("the blue one"), by pointing, or both.
- Such input combines three modalities: verbal, gestural, and graphical (the display). A common reference system is needed which can combine information across these modalities.

Background



Create a phase line between $\langle x,y \rangle$ grid coordinates:

CommandTalk (*speech-only*) - “create a line from nine four three nine six one to nine five seven nine six eight and call it phase line green.”

QuickSet (*multimodal*) - “phase line blue” while drawing a line.

- Oviatt '98 demonstrates clear task performance and user preference for multimodal over speech interfaces
 - 10% faster task completion, 23% fewer words, 35% fewer task errors, 35% fewer spoken disfluencies
 - People have difficulty articulating spatial information. **48% location errors on maps.**

(Oviatt, S. L., A. DeAngeli, et al. (1998). Integration and synchronization of input modes during multimodal human-computer interaction. Proceedings of Conference on Human Factors in Computing Systems: CHI '97, ACM)

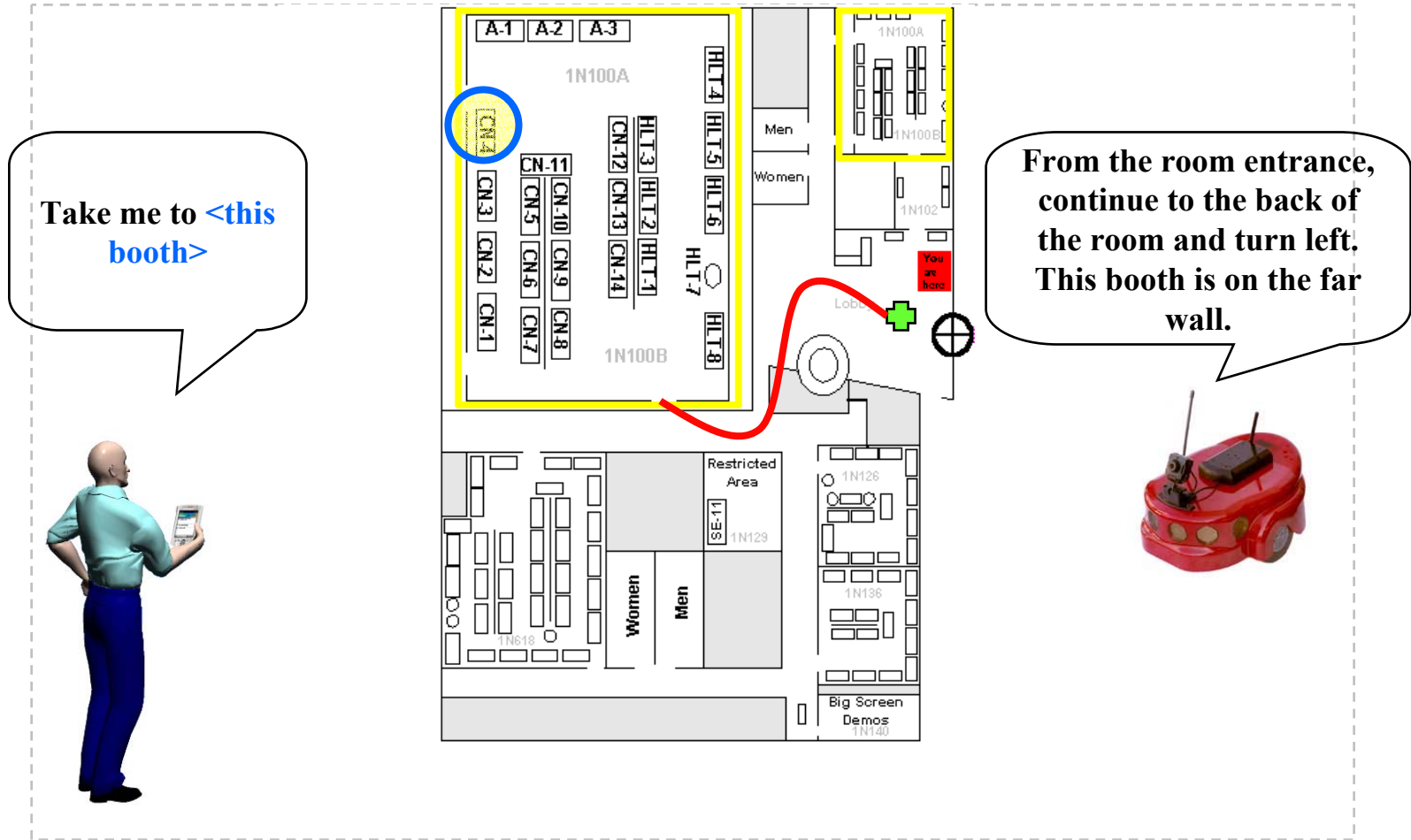
Objectives

- **Investigate and develop a framework for the interpretation of graphical, gestural, and language-based input for semantic understanding**
 - Demonstrate the power and expressiveness of multimodal input
 - Fuse information across modalities (display, gesture, speech) into a common discourse representation
- **This year we have two main objectives:**
 - Multi-threaded dialogue manager (DM) that supports input from several sources simultaneously
 - Demonstrate the use of PDA / speech to command a single robot in a simple navigational task

Activities

- **Re-engineering of TRINDIKIT architecture to support:**
 - Temporal requirements of multimodal processing
 - DARPA Communicator compliance
 - Robust, fault-tolerant operation
 - Consider support for multi-agent interaction
- **Proof-of-concept at MITRE Tech Symposium 2002 using both PDA interface and touch-based kiosk:**
 - Question and answer about the symposium
 - Robot control

Highlight



Highlight/Demonstration

- Implementation and demonstration of an open architecture conversational system at FY01 MITRE Tech Day

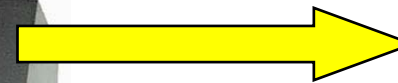
- Two-party
- Touch and speech
- Mixed-initiative human-machine conversation
- Question and answer task with life-like agent
- Domain-specific knowledge in a separate partition from the general purpose dialogue manager

Information Kiosk

Input:

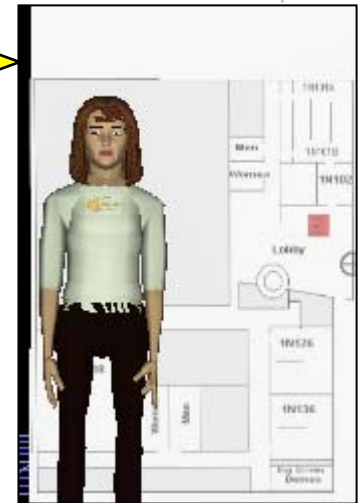
Built-in mike (no headset), touchscreen

“Mia, where are the big screen demos?”



Output:

Animated agent with synthesized speech, pointing to map of tech symposium



Impacts

- **Scalable user interfaces on large screen display, desktop, handheld and wearable devices**
- **Flexible component technology that may be quickly applied to sponsor-specific applications**
 - **Simulations**
 - **Robotics**
 - **Geographic and map-based systems**
- **Corpus standards**
 - **Trade survey and methodological considerations for gesture annotation**

Future Plans

