Collaborating with a Mobile Robot: An Augmented Reality Multimodal Interface

Scott A. Green*, **, XioaQi Chen*, Mark Billinghurst**
J. Geoffrey Chase*

*Department of Mechanical Engineering, University of Canterbury, Christchurch, NZ
(scott.green, xiaoqi.chen, geoff.chase@canterbury.ac.nz).
**Human Interface Technology Laboratory New Zealand (HIT Lab NZ), University of Canterbury
Christchurch NZ (mark.billinghurst@canterbury.ac.nz)

Abstract: We have created an infrastructure that allows a human to collaborate in a natural manner with a robotic system. In this paper we describe our system and its implementation with a mobile robot. In our prototype the human communicates with the mobile robot using natural speech and gestures, for example, by selecting a point in 3D space and saying "go here" or "go behind that". The robot responds using speech so the human is able to understand its intentions and beliefs. Augmented Reality (AR) technology is used to facilitate natural use of gestures and provide a common 3D spatial reference for both the robot and human, thus providing a means for grounding of communication and maintaining spatial awareness. This paper first discusses related work then gives a brief overview of AR and its capabilities. The architectural design we have developed is outlined and then a case study is discussed.

1. INTRODUCTION

In the future it will be more common for humans and robots to collaborate together. However, an effective system for human-robot collaboration must allow the human to communicate with the robot in a natural manner. The system we have developed allows for such communication through natural speech and gesture. We have integrated a dialogue manager and collaborative knowledge base that enables natural two-way communication spoken dialogue.

In a collaborative team effort it is important to capitalize on the strengths of each team member. For example, humans are good at problem solving and dealing with unexpected events while robots are good at repeated physical tasks and working in hazardous environments. Our system enables the human and robot to discuss a plan, after agreement between he robot and human, the robot then executes the plan. If an unexpected situation arises, the robot can discuss possible solutions with the human and arrive at a solution agreeable to both. This scenario is similar to the way a human team would collaborate.

Augmented Reality (AR) is a technology that overlays 3D virtual graphics onto the users view of the real world (Azuma, Baillot et al. 2001). AR allows real time interaction with these 3D graphics, enabling the user to reach into the augmented world and manipulate it directly. In human-robot collaborative endeavours the lack of situational awareness deteriorates robotic performance (Murphy 2004; Yanco, Drury et al. 2004). In our work we use AR to provide a common 3D graphic of the robot's workspace that both the human and robot can reference. In this way we enable the human to maintain situational awareness of the robot and its surroundings and give the human-robot team the ability to ground their communication (Clark and Brennan 1991).

The human can use natural gestures to communicate with the robot. The gesture processing is modal in that it allows for the use of gestures as commands, such as indicating go forward or turn, and also allows for gestures to select a point in 3D space coupled with spatial language such as "go here" or "go behind that". By coupling AR with spoken dialogue we have developed a multimodal interface that enables natural and efficient communication between the human and robot team members, thus enabling effective collaboration.

2. RELATED WORK

Bolt's work "Put-That-There" (Bolt 1980) showed that gestures combined with natural speech (multimodal interaction) lead to a powerful and more natural man-machine interface. Milgram *et al.* (Milgram, Zhai et al. 1993) highlighted the need for combining the attributes that humans are good at with those that robots are good at to produce an optimised human-robot team. Milgram *et al.* also pointed out the need for Human-Robot Interaction (HRI) systems that can transfer the interaction mechanisms that are natural for human communication to the precision required for machine information. Their approach used augmented reality overlays in a fixed work environment to enable the human 'director' to use spatial referencing to interactively plan and optimise the path of a robotic manipulator arm.

Skubic *et al.* (Skubic, Perzanowski et al. 2004) conducted a study on human-robot spatial dialogue. A multimodal interface was used, with input from speech, gestures, sensors and personal electronic devices. The robot was able to use dynamic levels of autonomy to reassess its spatial situation in the environment through the use of sensor readings and an evidence grid map. The result was natural human-robot spatial dialog enabling the robot to communicate obstacle locations relative to itself and receive verbal commands to move to an object it had detected.

Collaborative control was developed by Fong *et al.* (Fong, Thorpe et al. 2003) for mobile autonomous robots. The robots work autonomously until they run into a problem they are unable to solve. At this point, the robots ask the remote operator for assistance, allowing human-robot interaction and autonomy to vary as needed. Robot performance increases with the addition of human skills, perception and cognition, and benefits from human advice and expertise. The human and robots engage in dialogue (through messaging, not spoken dialogue), exchange information, ask questions and resolve differences.

In more recent work, Fong *et al.* (Fong, Kunz et al. 2006) note that for humans and robots to work together as peers, the system must provide mechanisms for these peers to communicate effectively. The Human-Robot Interaction Operating System (HRI/OS) introduced enables a team of humans and robots to work together on tasks that are well defined and narrow in scope. The agents are able to use dialogue to communicate and the autonomous agents are able to use spatial reasoning to interpret 'left of' type dialogue elements. The ambiguities arising from such dialogue are resolved through modelling the situation in a simulation.

Giesler *et al.* (Giesler, Salb et al. 2004) implemented an AR system that creates a path for a mobile robot to follow using voice commands and a 'magic wand' made from AR fiducial markers. Pointing the wand at the floor, which is calibrated using multiple fiducial markers, voice commands can be used to create nodes along a motion path. These nodes can be interactively moved or deleted. As goal nodes are reached, the node depicted in AR changes colour to keep the user informed of the robots progress. The robot will retrace steps if an obstruction is encountered and create a new plan to arrive at the goal destination.

Maida *et al.* (Maida, Bowen et al. 2006) showed through user studies that the use of AR resulted in significant improvements in robotic control performance. Similarly, Drury *et al.* (Drury, Richer et al. 2006) found that for operation of Unmanned Aerial Vehicles (UAVs) augmenting real-time video with pre-loaded terrain data resulted in significantly improved understanding of 3D spatial relationships compared to 2D video alone. The AR interface provided better situational awareness of the activities of the UAV. AR has also been used to display robot sensor information on the view of the real world (Collett and MacDonald 2006).

Our research is novel in that it uses AR to provide the remote user with a sense of presence in the robots workspace. AR enables the user to select a point in 3D space and refer to it using deictic references such as "here" and "there" and enables the use of prepositions such as "behind" combined with a gestural input to identify an object referred to as "this". A heads up display in the AR view shows the human the internal state of the robot. The intended motion of the robot is displayed in the AR scene prior to execution of the task. In this manner the robot and human discuss task execution and resolve differences and misunderstandings before the task is undertaken. Our interface also allows for the exchange of spoken dialog that can be initiated by any member of the team and combines this spatial language with gestures for natural communication.

3. AUGMENTED REALITY

Augmented Reality is a technology that overlays computer graphics onto the view of the real world of the user in real time. AR differs from virtual reality (VR) in that in a virtual environment the entire physical world is replaced by computer graphics. AR enhances rather than replaces reality. Azuma *et al.* (Azuma, Baillot et al. 2001) identify the following three characteristics of an AR interface:

- An AR interface combines real and virtual objects
- The virtual objects appear registered on the real world
- The virtual objects can be interacted with in real time

In a typical AR interface a user wears a head mounted display (HMD) with a camera mounted on it. This camera provides a view of the real world from the user's point of view. The camera is placed near the eyes of the user, as shown in Fig. 1. The output from the camera is fed into a computer and then into the HMD so the user sees the real world through the video provided by the camera.

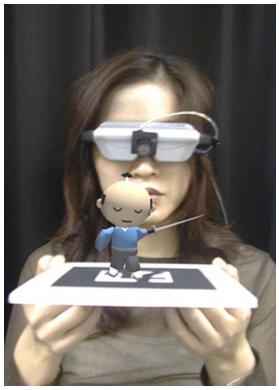

Fig. 1. AR interface with head mounted display, camera in its center, a fiducial marker and registered virtual image on the marker.

A collection of marked cards is placed in the real world with square fiducial patterns on them and a unique symbol in the middle of the pattern. Computer vision techniques provided by the ARToolKit library (ARToolKit 2007) are used to identify the unique symbol, calculate the camera position and orientation, and display 3D virtual images aligned with the position of the markers, see Fig. 2. In this manner the virtual images are seamlessly blended with the real world. The use of AR enables a user to experience a tangible user interface. Physical objects in the real world are manipulated to affect change in the 3D virtual scene (Billinghurst, Grasset et al. 2005).

Fig. 2. ARToolKit tracks a fiducial marker and aligns an object in AR that appears registered in the real world.

AR is an ideal platform for human-robot collaboration as it provides the following (Green, Billinghurst et al. 2007):

- The ability to enhance reality
- Seamless interaction between real and virtual environments
- The ability to share remote views
- The ability to visualize the robot relative to the task space
- Display of visual cues of robot's intentions and internal state
- Spatial cues for local and remote collaboration
- Support for tangible interface
- Support for use of deictic gestures and spatial language

AR provides a 3D view of the robot's work environment with the robot in it, which enables the user to maintain awareness of the robot relative to its workspace. The human uses the 3D visuals to reference locations in the robot's world. The system then easily relays this location information in the reference frame of the robot or human, whichever is appropriate. This ability to disambiguate reference frames enables the system to effectively ground communication.

4. ARCHITECTURE

A multimodal approach has been taken that combines speech and gesture through the use of AR that allows humans to naturally communicate with our mobile robot. Through this architecture the robot receives the discrete information to operate while allowing the human to communicate in a natural and effective manner by referencing objects, positions and intentions through natural speech and gesture. The human and robot maintain situational awareness by referencing the shared 3D visuals of the workspace in the AR environment.

The architectural design is shown in Fig. 3. The speech-processing module recognizes human speech and parses this speech into dialogue components. When a defined dialogue goal is achieved the required information is sent to the Multimodal Communication Processor (MCP). The speech-

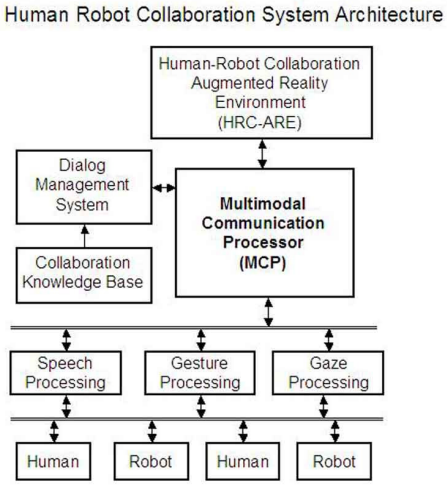

Fig. 3. The Human-Robot Collaboration system architecture.

processing module is also responsible for taking information from the MCP and robot and synthesizing this information into speech to enable effective dialogue with the human. The speech processing module is built on the Microsoft Speech Sapi 5 (MicrosoftSpeech 2007) .

Gesture processing enables the human to use deictic referencing and natural gestures to communicate effectively. The gesture-processing module recognizes gestures and passes this information to the MCP. The MCP combines the speech from the speech-processing module, the gesture information and uses the Human-Robot Collaboration Augmented Reality Environment (HRC-ARE) to effectively resolve ambiguous deictic references such as "here", "there", "this" and "that". The HRC-ARE also allows for the use of spatial references such as "behind this" and "on the right side of that". The human uses a real world paddle with fiducial markers attached to it to interact with the 3D virtual content.

The gesture processing is modal. A verbal command tells the system to process gestures with the paddle being a pointer or indicates to the system that natural gestures will be used. We have defined natural gestures from those used by participants in a WOZ study we ran to determine what kind of natural speech and gestures would be used to collaborate with a mobile robot (Green, Richardson et al. 2008). The user decides which type of gesture interaction to use. Natural gestures have been defined to communicate to the robot to move forward, turn at a relative angle, back up and stop. At any time the user can give a verbal command resulting in a true multimodal experience.

The paddle has a fiducial marker on the end opposite the handle. The paddle is flat and therefore has a fiducial marker on both sides, so that no matter which way the user holds the paddle the fiducial marker can be seen by the vision system. In the pointer mode a virtual pointer is attached to the paddle. When the paddle is used for natural gestures the virtual pointer does not appear. Instead different visual indicators appear to let the user know what command they are giving. If the user points the wand straight out in front of them it is

interpreted as a go forward gesture and an icon appears alerting the user of this.

When the paddle is moved to either side of straight in front of the user the system calculates the angle from straight ahead and converts this information into a turn. To turn the robot in place the user starts from the straight up position and rotates their arm about their elbow to the right or left. The severity of the turn the robot makes is proportional to the amount the user rotates their arm. To go in the reverse direction the user places the paddle in a straight up position. Any position of the paddle not specifically defined is interpreted as a stop command and is relayed to the user by displaying a stop sign. See Fig 4. for various paddle-gesture commands.

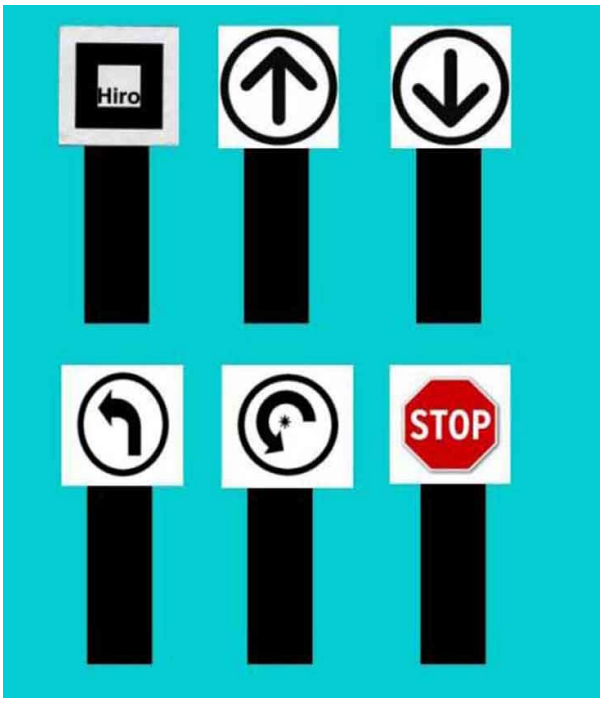

Fig. 4. Paddle with fiducial marker (top left) and augmented graphics to indicate mode paddle is in.

The gaze-processing module defines the gaze direction of the user through the use of the ARToolKit and tracking of the fiducial markers. The gaze direction of the user in the AR environment is used to define spatial terms such as "behind" and "to the right of". By knowing where the user is in reference to the objects in the virtual scene spatial references can be defined in the reference frame of the user, as described in (Irawati, Green et al. 2006). This information is easily translated into the reference frame of the robot since the HRC-ARE knows the location of the robot and all the virtual objects. The desired location is then sent to the robot where it uses its autonomous capabilities to move to the position in the real world.

The Dialogue Management System (DMS) is aware of the communication between the human and robot. The MCP takes the information from the speech, gesture and gaze processing modules together with the information generated from the HRC-ARE and supplies it to the DMS. The DMS is responsible for combining this information and comparing it to the information stored in the Collaboration Knowledge Base (CKB). The CKB contains information pertaining to what is needed to complete the desired tasks that the human-robot team wishes to complete. The DMS then responds through the MCP to either the human team member or the robot facilitating dialogue and tracking when a command or request is complete.

The MCP is responsible for receiving information from the other modules in the system and sending information to the appropriate modules. The MCP is thus responsible for combining multimodal input, registering this input into something the system can understand and then sending the required information to other system modules for action. The effect of this system design is that the human is able to use natural speech and gestures to collaborate with the robot.

5. CASE STUDY

As a case study we used a Lego Mindstorms NXT (Lego 2007) mobile robot in the Tribot configuration to collaborate with (see Fig. 5). To incorporate the mobile robot into our system we used NXT++ (NXT++ 2007), an interface to the Mindstorms robot written in C++. We chose to use a Lego Mindstorms robot because it is a simple platform to prove out the functionality of our human-robot collaborative system.

Fig. 5. Lego Mindstorms NXT robot in the Tribot configuration.

The case study task was to have a human collaborate with the robot to navigate a maze, as shown in Fig. 6. A desired path was defined and various obstacles were placed in this path that the robot would have to maneuver around. The robot was unaware of the path plan and had to collaborate with the human to get through the defined path.

Our robot had only one ultrasonic sensor on the front to sense objects and measure the distance to them. It also had a touch sensor on the front that would stop the robot if triggered to avoid colliding with something. The limited sensing ability of the robot allowed us to take advantage of dialogue to ensure the robot took a safe path. An example would be when the robot had to back up. With no rear sensors the robot was unable to determine if a collision was imminent. In this case the robot asked the human if it was ok to move in reverse without hitting anything. Once the robot received confirmation the path was clear, it began movement. Since

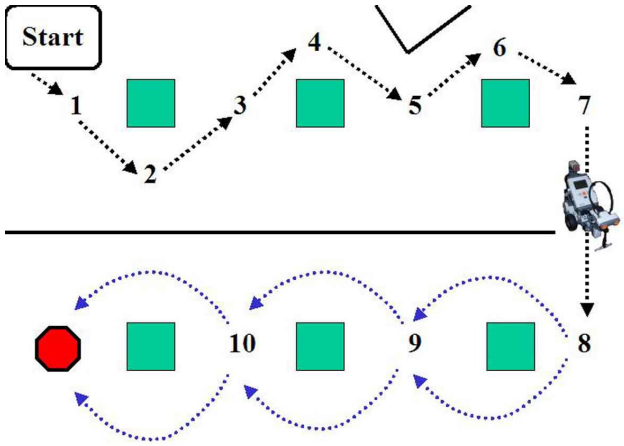

Fig. 6. Maze for case study, black lines indicate defined path, blue lines indicate users choice.

the robot had to ask for guidance, the user was aware that the robot might need assistance in completing the maneuver.

The robot's environment was modelled in 3D and used as the virtual scene in AR. This set-up gave the human a feeling of presence in the robot's world. The system allows the human to naturally communicate with the robot in the modality most comfortable to the user. Given the restrictions of our Mindstorms robot sensors the human had to do more monitoring than would be necessary with a more autonomous robot.

A heads up display was used to keep the human informed of the internal state of the robot. The human could easily see the directions the robot was moving, the battery level, motor speeds, paddle mode and server status. Fig. 7 is an example of the human view through the HMD. The robots internal state is easily identifiable as is the robots intended path and progress.

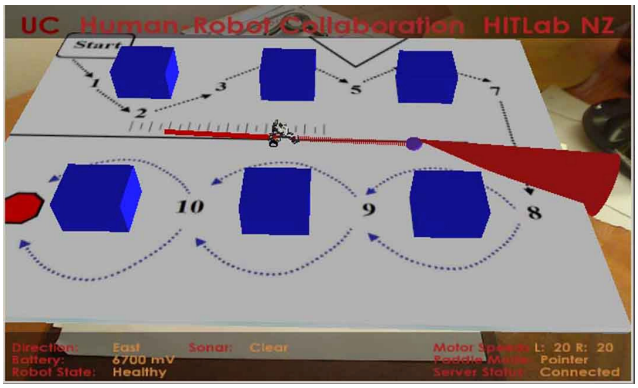

Fig 7. Robot state as seen by the human through the HMD.

The human sets the modality of the pointer with a verbal command. The pointer can be used to portray defined gestures for move forward, turn at an angle, stop and move backwards. Changing the modality of the pointer the user can select a point in 3D space and tell the robot to "go there". The user can also select an object and tell the robot to "go to the right of that" or "go behind this".

Because of the limited autonomy of the robot it used spoken dialogue when it was unsure if it could proceed without a collision. When a request was made for the robot to go behind something, the robot asked the human to which side it should go. The user was able to say "go to the right" which is interpreted as the right in the robot's reference frame. The user can also say "go to my right" and the system will use the knowledge of the position of the human, object and robot, distinguish what "go to my right" means to the robot and send the appropriate command to the robot. This disambiguation was made possible through the use of AR.

6. FULL-SCALE VALIDATION STUDIES

We are in the process of designing and running full-scale validation studies to determine the robustness and effectiveness of our human-robot collaboration system. The studies will highlight telepresence in the sense that the human collaborator will be located remotely from the robots with which the human will be interacting. The participants will use three modalities to interact with the system:

- Speech only interface
- Gesture only interface
- Multimodal: speech and gesture interface

Alternatively, or in combination with the different modalities, the users will have three ways to interact with the system:

- Head Mounted Display (HMD) AR system
- Non-HMD AR system, using screen display instead
- 2D mouse interaction

The studies will measure the following:

- Completion times
- Crashes
- Distance travelled
- Situational awareness
- Subjective measures of intuitiveness of interaction

7. CONCLUSIONS

In this paper we introduced our prototype system for human-robot collaboration. This system uses Augmented Reality to provide a means for a human to effectively communicate with a robot. AR provides a common 3D graphic of the robot's workspace that the human can interact with. This graphic is used as a reference for both the human and robot thus enabling robust grounding of communication. Our system allows the human to maintain situational awareness of the robot through the use of AR. The robot displays its internal state and intentions in the AR imagery.

We combined spatial language with natural gestures to achieve a multimodal interface. This interface enables the human to communicate in a natural manner using deictic gestures. AR disambiguates these deictic gestures and sends the robot information in a form that the robot needs to operate. The system is aware of the position of the team

members and objects thus allowing the use of different reference frames. In this manner our system enables a human to effectively collaborate with a mobile robot.

REFERENCES

ARToolKit (2007). http://www.hitl.washington.edu/artoolkit/, accessed August 2007

Azuma, R., Y. Baillot, et al. (2001). Recent advances in augmented reality, *IEEE Computer Graphics and Applications*, 21, (6), 34-47

Billinghurst, M., R. Grasset, et al. (2005). Designing Augmented Reality Interfaces, *Computer Graphics SIGGRAPH Quarterly*, 39(1), 17-22 Feb

Bolt, R. A. (1980). Put-That-There: Voice and Gesture at the Graphics Interface, *In Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, 14, 262-270

Clark, H. H. and S. E. Brennan (1991). Grounding in Communication, *Perspectives on Socially Shared Cognition*, L. Resnick, Levine J., Teasley, S., Washington D.C., American Psychological Association**:** 127 - 149

Collett, T. H. J. and B. A. MacDonald (2006). Developer Oriented Visualisation of a Robot Program, *Proceedings 2006 ACM Conference on Human-Robot Interaction, March 2-4*, 49-56

Drury, J., J. Richer, et al. (2006). Comparing Situation Awareness for Two Unmanned Aerial Vehicle Human Interface Approaches, *Proceedings IEEE International Workshop on Safety, Security and Rescue Robotics (SSRR). Gainsburg, MD, USA* August

Fong, T., C. Kunz, et al. (2006). The Human-Robot Interaction Operating System, *Proceedings of 2006 ACM Conference on Human-Robot Interaction, March 2-4*, 41-48

Fong, T., C. Thorpe, et al. (2003). Multi-robot remote driving with collaborative control, *IEEE Transactions on Industrial Electronics*, 50, (4), 699-704

Giesler, B., T. Salb, et al. (2004). Using augmented reality to interact with an autonomous mobile platform, *Proceedings-2004 IEEE International Conference on Robotics and Automation, Apr 26-May 1*, New Orleans, LA, United States, Institute of Electrical and Electronics Engineers Inc., Piscataway, United States

Green, S. A., M. Billinghurst, et al. (2007). Human-Robot Collaboration: An Augmented Reality Approach; A Literature Review and Analysis, *Proceedings of 3rd International Conference on Mechatronics and Embedded Systems and Applications (MESA 07), September 4-7*, Las Vegas Nevada

Green, S. A., S. M. Richardson, et al. (2008). Multimodal Metric Study for Human-Robot Collaboration, *Proceedings of 1st International Conference on Advances in Computer-Human Interaction (ACHI-08), February 10 - 15*, Sainte Luce, Martinique

Irawati, S., S. Green, et al. (2006). Move the Couch Where? Developing an Augmented Reality Multimodal Interface, *In Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2006)*, Santa Barbara, California

Lego (2007). http://mindstorms.lego.com/, accessed August 2007

Maida, J., C. Bowen, et al. (2006). Enhanced Lighting Techniques and Augmented Reality to Improve Human Task Performance, *NASA Tech Paper TP-2006-213724*, July

MicrosoftSpeech (2007). http://www.microsoft.com/speech/default.mspx, accessed August 2007

Milgram, P., S. Zhai, et al. (1993). Applications of Augmented Reality for Human-Robot Communication, *In Proceedings of IROS 93: International Conference on Intelligent Robots and Systems*, Yokohama, Japan

Murphy, R. R. (2004). Human-robot interaction in rescue robotics, *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34, (2), 138-153

NXT++ (2007). www.nxtpp.sourceforge.net/index.php, accessed August 2007

Skubic, M., D. Perzanowski, et al. (2004). Spatial language for human-robot dialogs, *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34, (2), 154-167

Yanco, H. A., J. L. Drury, et al. (2004). Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition, *Human-Computer Interaction Human-Robot Interaction*, 19, (1-2), 117-149