Multimodal HCI for Robot Control: Towards an Intelligent Robotic Assistant for People with Disabilities

Zunaid Kazi, Shoupu Chen, Matthew Beitler, Daniel Chester and Richard Foulds
Applied Science and Engineering Laboratories,
Unversity of Delaware/A.I. duPont Institute Wilmington,
DE 19899, USA
{kazi,chen,beitler,chester,foulds}@asel.udel.edu

Abstract

The Multimodal User Supervised Interface and Intelligent Control (MUSIIC) project is working towards the development of an assistive robotic system which integrates human-computer interaction with reactive planning techniques borrowed from artificial intelligence. The MUSIIC system is intended to operate in an unstructured environment, rather than in a structured workcell, allowing users with physical disabilities considerable freedom and flexibility in terms of control and operating ease. This paper reports on the current status of the MUSIIC project.

Introduction

One of the most challenging problems in rehabilitation robotics has been the design and development of an efficient control mechanism that allows users with motor disabilities to manipulate their environment in an unstructured domain. Physical limitations of motion range, coordination of movement and grasping, and lack of strength all contribute to a decreased ability to perform normal manual tasks. Fortunately, in principle, this loss may be compensated for by the use of assistive robots which may act on the user's behalf in carrying out the manipulation.

The Multimodal User Supervised Interface and Intelligent Control (MUSIIC) project is developing an assistive robot system that uses a multimodal (speech and gesture) interface to allow people with disabilities to manipulate real world 3-D objects (Chen et al., 1994, Beitler et al., 1995b, Beitler et al., 1995a, Kazi et al., 1995a, Kazi et al., 1995b). The MUSIIC strategy is a novel approach for an intelligent assistive telerobotic system: speech-deictic gesture control is integrated with a knowledge-driven reactive planner and a stereo-vision system. The system is intended to meet the needs of individuals with physical disabilities and operate in an unstructured environment, rather than in a structured workcell allowing the user considerable freedom and flexibility in terms of control and operating ease.

The MUSIIC strategy utilizes a stereo-vision system to determine the three-dimensional shape and pose of objects and surfaces which are in the immediate environment, and provides an object-oriented knowledge base and planning system which superimposes information about common objects in the three-dimensional world. This approach allows the user to identify objects and tasks via a multimodal user interface which interprets their deictic gestures and speech inputs. The multimodal interface performs a critical disambiguation function by binding the spoken words to a locus in the physical work space. The spoken input is also used to supplant the need for general purpose object recognition. Instead, three-dimensional shape information is augmented by the user's spoken word which may also invoke the appropriate inheritance of object properties using the adopted hierarchical object-oriented representation scheme.

To understand the intricacies and embodied meaning of the numerous modal inputs, we have also designed a graphical simulation of the multimodal environment. This simulation will allow us to study and better understand the interplay between the user and the MUSIIC system. Additionally, the simulated environment will be an integral part of the actual MUSIIC system by providing the user a visualization which depicts the planner's interpretation of the information gathered by the system. The MUSIIC system's ability to determine the superquadric shape representation of the scene from stereo vision enables the graphical simulation to dynamically model a variety of real world entities and objects.

A very simple illustration (Figures 1 and 2) describes how our proposed system functions in a real-world scenario.

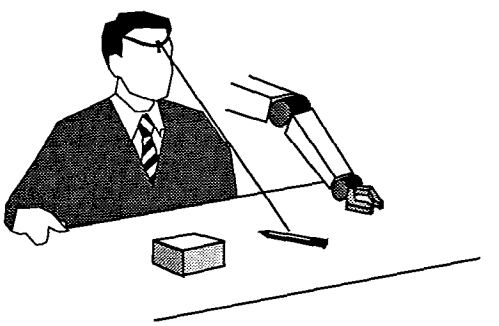

Figure 1: That's a pen

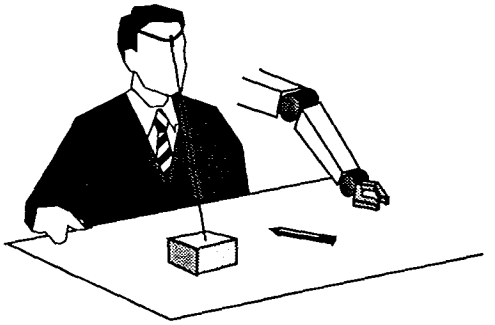

Figure 2. "Put the pen there"

The user approaches a table on which there are a *pen* and a *box* The user points to the pen, and says, *that's a pen.* The user points to the box and says *put the pen there*, indicating that the pen must be moved to the location *there.* The system then executes the user's intentions.

Justification

Rehabilitation robotics literature describes many demonstrations of the use of robotic devices by individuals with disabilities (Foulds, 1986,Bacon et al., 1994).

In general, these interface methods have taken two distinct approaches. In a command based approach, users rely on the activation of pre-programmed tasks (Fu, 1986, van der loos et al., 1990), while in a contrasting approach, the user directly controls the movement of the manipulator much like a prosthesis (Zeelenberg, 1986, Kwee, 1986). The limitations of a command based interface were discussed by Michalowski et al., 1987. The effectiveness of the command system is limited by the need for a reasonably structured environment and a limited number of commands. While direct control allows the user to operate in an unstructured environment, physical as well as cognitive loads on the user precludes the development of an efficient and useful assistive robot. At the other extreme are completely autonomous systems that perform with effectively no user supervision, the long elusive goal of AI, robotics and machine vision communities. Unfortunately, this goal seems far away from the state of the art at this point, although many important incremental advances have been forthcoming in the past decades. Furthermore, absolute automation poses a set of problems stemming from incomplete *a priori* knowledge about the environment, hazards, strategies of exploration, insufficient sensory information, inherent inaccuracy in the robotic devices and the mode of operation (Sheridan, 1992).

What one should strive for is a synergistic integration of the best abilities of both *"humans"* and *"machines"*. Humans excel in creativity, use of heuristics, flexibility and *common sense*, whereas machines excel in speed of computation, mechanical power and ability to persevere. While progress is being made in robotics in areas such as machine vision and sensor based control, there is much work that needs to be done in high level cognition and planning. We claim that the symbiosis of the high level cognitive abilities of the human, such as object recognition, high level planning, and event driven reactivity with the native skills of a robot can result in a human-robot system that will function better than both traditional robotic assistive systems and autonomous systems. We describe a system that can exploit the low-level machine perceptual and motor skills and excellent AI planning tools currently achievable, while allowing the user to concentrate on handling the problems that users are best suited for, namely high-level problem solving, object recognition, error handling and error recovery. By doing so, the cognitive load on the user is decreased, the system becomes more flexible, less fatiguing, and is ultimately a more effective assistant.

Our multimodal interface mechanism would allow the user to remain in the loop, while lessening the physical demands. By utilizing a multimodal interface to combine input evidence from a user dialogue, perceptual and planning requirements of the system can be relaxed to the point where existing semi-autonomous techniques are sufficient to carry out tasks and make the system practical. By engaging in dialogue with the user in such a way that natural deictic gestures and voice input can be used to carry out a task, the system gains many of the advantages present in direct manipulation interfaces. The user can directly designate objects and locations in the environment around him/her, and use natural language to describe the desired actions on those objects and locations. By combining different modalities, rather than attempting to constrain dialogue to one modality, great simplification of processing can be accomplished, as has been demonstrated by several multimodal systems that have been developed for graphical user interfaces (Bolt, 1980, Koons, 1994). This simplified processing allows for less delay in the processing of user interaction, which supports faster system response to user actions, which improves user task completion times and to result in less frustration (Shneiderman, 1992).

MUSIIC Architecture

In this section we discuss both the implementation as well as the architecture of the MUSIIC system.

System Description

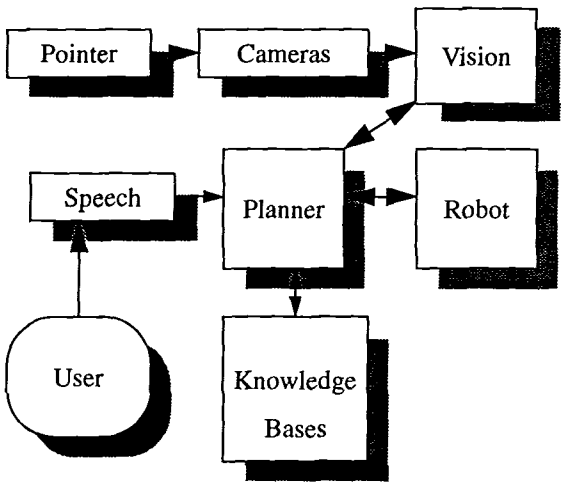

Figure 3: Components

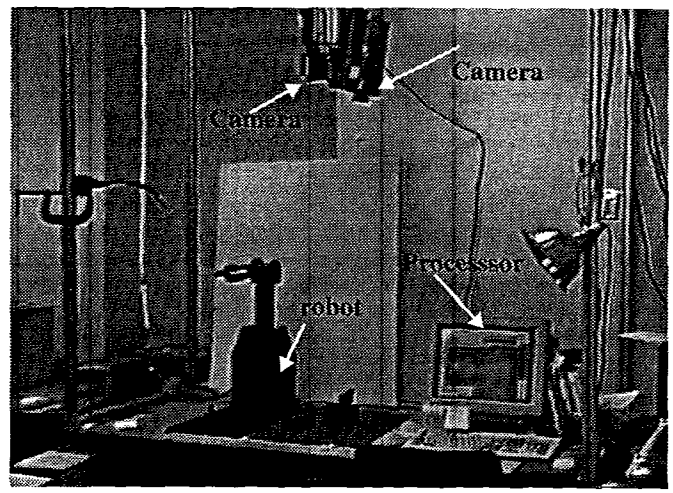

Figure 4: Physical set-up

The previous sections lead naturally to a description of the essential components of the MUSIIC system (Figure 3). We require a ***planner*** that will interpret and satisfy user intentions. The planner is built upon ***object oriented knowledge bases*** that allow the users to manipulate objects that are either known or unknown to the system. A ***speech input*** system is needed for user inputs, and a ***gesture identification*** mechanism is necessary to obtain the user's deictic gesture inputs. An ***active stereo-vision*** system is necessary to provide a snap-shot of the domain; it returns object shapes, poses and location information without performing any object recognition. The vision system is also used to identify the focus of the user's deictic gesture, currently implemented by a laser light pointer, returning information about either an object or a location. The planner extracts user intentions from the combined speech and gesture input. It then develops a plan for execution on the world model built up from the *a priori* information contained in the knowledge bases, the real-time information obtained from the vision system, the sensory information obtained from the robot arm, as well as information previously extracted from the user dialog. Prior to execution, the system allows the user to preview and validate the planner's interpretation of the user's intentions via a 3-D graphically ***simulated environment*** (Beitler et al., 1995a). Figure 4 shows the actual system set-up

The high level planner

The high-level planner is described briefly in this section Details of the planning mechanism can be found in Kazi et al., 1995a. Our architecture for task planning incorporates a novel reactive planning system where the user is an integral component of the planning mechanism. The planning mechanism is based on an object-oriented knowledge base and an object oriented plan library.

Our hierarchical human-machine interface and object oriented representation allows the user to interact with the planning system at any level of the planning hierarchy, from low-level motion and grasp planning to high-level task planning of complex tasks such as feeding. The generic plans and specialized plans are supplemented by user interaction whenever incomplete information precludes the development of correct plans by taking over control of the planning mechanism or providing information to the knowledge bases to facilitate the development of a plan capable of handling a new or uncertain situation. Furthermore, incomplete sensory information may be supplemented by user input, enabling the planner to develop plans from its plan library without the need for extensive user intervention.

Given this underlying architecture, the system first determines what the user wants, and then makes plans to accomplish the task. As a consequence of insufficient information, uncertainty, advent of new information, or failure of a plan, the system engages in a dialogue with the user which enables the planner to revise its plans and actions.

Vision

For our multimodal system, the vision requirement is to provide the knowledge based planning system with parameterized shape and pose information of the objects in the immediate environment. This information can then be used to fill slots in the object oriented representation and support both the system planning and simulation activities. The vision processing proceeds in three phases: extraction of highly precise 3-D point information using a calibrated line-based stereo

matching algorithm, segmentation of the point sets into object-based sets, and non-linear minimization to fit parameterized shapes to respective objects in the scene, details of which can be found in Chen et al., 1994.

Simulated environment

We are developing a simulation environment that will allow us to investigate, in a low risk fashion, the use of the multiple modalities of the user to control a rehabilitation robot (Beitler et al., 1995b). The type of simulation we are using has been referred to as a "fish-tank" environment, in which the individual feels that he is on the outside looking in through the side of a fish-tank (monitor screen) (Ware & Jessome, 1988). This simulation models not only the robot and the domain but also the interplay between user intentions and the robot's perception of these intentions. This simulation mechanism has been developed using JACK (Badler et al., 1993).

The multimodal interface

Researchers have proposed a number of systems which investigate alternate modes of human-computer interaction in addition to speech and vision based ones. Work has been carried out in using gestures and hand pointing as a mode of man-machine interface. In some systems, researchers have required the users to use hand gloves (Cipolla et al., 1992, Fukimoto et al., 1992), while others require calibration for each individual's hand shapes and gestures (Wiemer & Ganapathy, 1989). Cipolla et al. report preliminary work on a gesture-based interface for robot control (Cipolla et al., 1992). Their system requires no physical contact with the operator, but uses un-calibrated stereo-vision with active contours to track the position and pointing direction of a hand. Pook describes a deictic gesture based tele-assistance system for direct control of a telerobot, although the system lacks a perceptual component (Pook, 1994). Funda et al. describe a teleprogramming approach which extracts user intentions from interaction with a virtual model of a remote environment, but their system requires an *a priori* 3-D model of the remote scene (Funda et al., 1992).

Work is also being done in attempting to extend this concept by using multiple modes of human-machine interfacing. (Bolt, 1980, Cannon, 1992, Cannon et al., 1994) MUSIIC extends the combined deictic gesture and spoken word of Bolt to true 3-D environments manipulated by a robot. The gesture control and the spoken input are used to make a general purpose object recognition module unnecessary. Instead, 3-D shape information is augmented by the user's spoken word which may also invoke the appropriate inheritance of object properties using the adopted hierarchical object-oriented representation scheme.

In the introduction we argued how using a multimodal interface to combine input evidence from a user dialogue mitigates the requirements for perceptual and planning systems to support direct manipulation. In the following sections we discuss the multimodal control input language.

Semantic interpretation for robot control

In order to devise a practical command input interpretation mechanism we restricted both the nature of our speech input as well as our gesture input.

Speech

Consider the user command:

```
Put the book on the table
```

While a fully fledged natural language system combined with a state-of-the-art gesture recognition mechanism may allow the user more expressive power, the state-of-the-art in these two areas makes this a distant goal. At the same time, the requirements of the domain place some constraints on the choice of modalities and the degree of freedom in expressing user intentions. A multimodal combination of speech and pointing is a better alternative for use as an assistive device, where the input speech is a restrictive sub-set of natural language, a pseudo-natural language (PNL). We then can apply model-based procedural semantics (Crangle et al., 1988), where words are interpreted as procedures that operate on the model of the robot's physical environment. One of the major questions in procedural semantics has been the choice of candidate procedures. Without any constraints, no procedural account will be preferred over another and there will not be any shortage of candidate procedures. The restrictive PNL and the finite set of manipulatable objects in the robots domain provide this much needed set of constraints.

Gesture

Similarly, the needs of users with disabilities also restrict the choice of gestures. Our gesture of choice is deictic gesture, which is simply pointing. In the general case, not only does pointing have the obvious function of indicating objects and events in the real world, it also plays a role in focusing on events/objects/actions that may not be objectively present (McNeill, 1982). The choice of deictic gestures allows us to use any number of devices, not restricted to the hand, to identify the user's focus. While our research is investigating the use of a laser pointer to identify the user's focus of intentions, any device that is able to indicate a domain object can be used, such as eye tracking systems, mouse on a control panel, etc.

Combining speech and gesture

Like natural languages, gestures convey meanings. While their expressiveness is not inferior to natural languages, the methods used by gestures are fundamentally different from that of language. Segmentation and linearization to form a hierarchically structured string of words that are the essential feature of a linguistic system is based on the fact that language can vary only along the temporal dimension. Gestures are different in every way. McNeill describes a number of ways in which gestures are different (McNeill, 1982).

- Gestures are global-synthetic
- Gestures are combinatoric
- Gestures have no standards of form
- Gestures have no duality of patterns

These inherent differences makes gesture identification a very difficult task. However, while gestures and speech differ from each other in a number of fundamental ways, they are also closely linked in many ways.

- Gestures occur during speech
- Gestures and speech are semantically and pragmatically co-expressive
- Gestures and speech are synchronous

Restricting our choice of gestures to pointing gestures only allows us to use the above properties to extract user intentions in an unambiguous way. We are using pointing gestures to identify the user's focus of attention, to indicate either an object or a location. Currently, speech deictics "that" and "there" are being used in conjunction with pointing to identify the user's focus. The interpretation process must to capture the user's actions in speech and gesture within the domain of operation and then attempt to match them to elements in the system's domain knowledge base. We are able to extract the combined user intention by the use of time-stamps that allow us to identify which object or which location was the focus of intention during the user's deictic utterances. Each word is tagged with a time stamp, and the vision system is continuously scanning the world and storing a history of points identified by the gesture (in our case the laser pointer). Depending upon whether the speech deictic was a "that" or a "there", the procedures encoded with each word then returns either an object or location respectively. The required action is then invoked upon the returned values.

Result and Illustration

The current operational implementation of MUSIIC is able to manipulate objects of generic shapes at arbitrary locations. A rudimentary knowledge base of objects has been constructed which contains several hierarchies of abstract objects as well as a small set of object instantiations such as "*box*", "*pencil*", "*cup*", "*straw*" and "*book*".

A set of robot control primitives are used to build up higher level task commands with which the user instructs the assistive robot. The robot primitives include approaching, grasping and moving an object amongst others. The vision system first takes a snapshot of the domain and returns to the planner object sizes, shapes and locations. This information is then combined with the knowledge base of objects to model the workspace in question. The user then points to objects using a laser light pointer while verbally instructing the robot to manipulate an object.

For example, the user may say "`Put that here`", while pointing at an object as she says "`that`" and pointing to a location as she says "`here`". First, the combined gesture and verbal deictic is interpreted by the planner based on information extracted from the vision system as well as the object knowledge base. The planner then uses the plan knowledge base to approach and grasp the object and then move the object to the desired location.

In addition to high level commands as illustrated above, the user is also able to instruct the robot at a lower level, by commands such as "`move there`", "`open gripper`", "`move down`", "`close gripper`", "`move here`" to obtain the same functionality as the "`move that here`" instruction.

Several scripts are shown below to describe what the MUSIIC system is currently capable of handling.

(Scene 1)

Domain: Boxes of different colors
User points to a green box and says "*that's a box*"
User then says "*It's green*"
She then instructs the system to "*Put the green box on top of the blue box*"
The robot arm approaches and picks up the green box and deposits it on top of the blue box.

(Scene 2)

Domain: A cup and a straw
User points to the straw and says "*that's a straw*"
User points to the cup and says "*that's a cup*"
User then points to the cup and says "*insert the straw into the cup*"
The robot arm approaches and picks up the straw. It then orients the straw in the

correct orientation and inserts the straw into the cup.

Discussion

While MUSIIC is still very much a work in progress, the current test-bed implementation has amply demonstrated the flexibility in use of an assistive robot achievable by our multimodal RUI built on top of an intelligent planner. Work continues to flesh out the complete object hierarchy that will allow the planner to plan tasks at any level of specialization, from objects about which nothing is known except what the vision system returns, to objects which are well known, such as a cup often used by the user. The reactive component is also nearing completion. Reactivity will be achieved in two ways: An autonomous run-time reactivity will be obtained through sensor fusion, and a human centered reactivity will be used where the user can take over the planning process when the planner fails to make correct plans as a consequence of incomplete information or catastrophic failures. The user will engage in a dialog with the system, either to update the knowledge bases or to perform plan correction or editing.

Conclusion

Human intervention as well as an intelligent planning mechanism are essential features of a practical assistive robotic system. We believe our multimodal robot interface is not only an intuitive interface for interaction with a three-dimensional unstructured world, but it also allows the human-machine synergy that is necessary for practical manipulation in a real world environment. Our novel approach of gesture-speech based human-machine interfacing enables our system to make realistic plans in a domain where we have to deal with uncertainty and incomplete information.

Acknowledgment

Work on this project is supported by the Rehabilitation Engineering Research Center on Rehabilitation Robotics, National Institute on Disabilities and Rehabilitation Research Grant #H133E30013 (Department of Education), Rehabilitation Services Administration Grant #H129E20006 and Nemours Research Programs.

References

Bacon, D. C., Rahman, T., & Harwin, W. S., Eds. 1994. *Fourth International Conference on Rehabilation Robotics*, A. I. DuPont Institute, Wilmington, Delaware, USA. Applied Science and Engineering Laboratories.

Badler, N. I., Phillips, C. B., & Webber, B. L. 1993. *Simulating humans*. Oxford University Press.

Beitler, M., Foulds, R., Kazi, Z., Chester, D., Chen, S., & Salganicoff, M. 1995a. A simulated environment of a multimodal user interface for a robot. In *RESNA 1995 Annual Conference* (pp. 490–492). Vancouver, Canada: RESNA press.

Beitler, M., Kazi, Z., Salganicoff, M., Foulds, R., Chen, S., & Chester, D. 1995b. Multimodal user supervised interface and intelligent control (MUSIIC). In *AAAI 1995 Fall Symposium Series on Embodied Language and Action* MIT, Cambridge, Massachussetts.

Bolt, R. A. 1980. *Computer Graphics, 14*(3), 262–270.

Cannon, D. 1992. *Point and direct telerobotics: Object level strategic supervision in unstructured human-machine interface.* Unpublished doctoral dissertation, Stanford University, Department of Mechanical Engineering.

Cannon, D., Thomas, G., Wang, C., & Kesavadas, T. 1994. Virtual reality based point-and-direct robotic system with instrumented glove. *International Journal of Industrial Engineering – Applications and Practice, 1*(2), 139–148.

Chen, S., Kazi, Z., Foulds, R., & Chester, D. 1994. Multimodal direction of a robot by individuals with a significant disabilities. In *Second International Conference on Rehabilitation Robotics 1994* (pp. 55–64).

Cipolla, R., Okamotot, Y., & Kuno, Y. 1992. Qualitative visual interpretation of hand gestures using motion parallax. In *IAPR Workshop on Machine Vision Applications* (pp. 477–482).

Crangle, C., Liang, L., Suppes, P., & Barlow, M. 1988. Using Englishto instruct a robotic aid: An experiment in an office-like environment. In *Proc. of the International Conf. of the Association for the Advancement of Rehabilitation Technology* (pp. 466–467).

Foulds, R. A. 1986. *Interactive robotics aids-one option for independent living: an international perspective.* World Rehabilitation Fund.

Fu, C. 1986. An independent vocational workstation for a quadriplegic. In R. Foulds (Ed.), *Interactive robotics aids-one option for independent living: an international perspective*, volume 1 of *37* (pp. 42). World Rehabilitation Fund.

Fukimoto, M., Mase, K., & Suenga, Y. 1992. A synthetic visual environment with hand gesturing and voice input. In *IAPR Workshop on MAchine Vision Applications* (pp. 473–476).

Funda, J., Lindsay, T. S., & Paul, R. P. 1992. Teleprogramming: Towards time-invariant telemanipulation. *Presence, 1*(1), 29–44.

Kazi, Z., Beitler, M., Salganicoff, M., Chen, S., Chester, D., & Foulds, R. 1995b. Intelligent telerobotic assistant for people with disabilities. In *SPIE's International Symposium on Intelligent Systems: Telemanipulator and Telepresence Technologies II*: SPIE.

Kazi, Z., Beitler, M., Salganicoff, M., Chen, S., Chester, D., & Foulds, R. 1995a. Multimodal user supervised interface and intelligent control (MUSIIC) for assistive robots. In *1995 IJCAI workshop on Developing AI Applications for the Dis-*

abled (pp. 47–58).
Koons, D. P. 1994. Capturing and interpreting multi-modal descriptions with multiple representations. In *Intelligent Multi-Media Multi-Modal Systems.*
Kwee, H. 1986. Spartacus and Manus: Telethesis developments in france and the netherlands. In R. Foulds (Ed.), *Interactive robotic aids-one option for independent living: An international perspective*, volume Monograph 37 (pp. 7–17). World Rehabilitation Fund.
McNeill, D. 1982. *Hand and mind : What gestures reveal about thought.* The University of Chicago Press.
Michalowski, S., Crangle, C., & Liang, L. 1987. Experimental study of a natural language interface to an instructable robotic aid for the severely disabled. *In Proc. of the 10th Annual Conf. on Rehabilitation Technology*, (pp. 466–467).
Pook, P. 1994. Teleassistance: Contextual guidance for autonomous manipulation. In *National Conference on Artificial Intelligence*, volume 2 (pp. 1291–1296). Menlo Park, CA: AAAI.
Sheridan, T. B. 1992. *Telerobotics, automation, and human supervisory control.* Cambridge, MA: The MIT Press.
Shneiderman, B. 1992. *Designing the user interface : strategies for effective human-computer interaction.* Addison-Wesley.
van der loos, M., Hammel, J., Lees, D., Chang, D., & Schwant, D. 1990. *Design of a vocational assistant robot workstation.* Annual report, Rehabilitation Research and Development Center, Palo Alto VA Medical Center, Palo Alto, CA.
Ware, C. & Jessome, R. 198). Using the bat: A six dimensional mouse for object placement. *IEEE Computer Graphics and Applications*, *8*(6), 155–160.
Wiemer, D. & Ganapathy, S. G. 1989. A synthetic visual environment with hand gesturing and voice input. In *CHI* (pp. 235–240).
Zeelenberg, A. 1986. Domestic use of a training robot-manipulator by children with muscular dystrophy. In R. Foulds (Ed.), *Interactive robotic aids-one option for independent living: An international perspective*, volume Monograph 37 (pp. 29–33). World Rehabilitation Fund.