Smart Sight: A Tourist Assistant System

Jie Yang, Weiyi Yang, Matthias Denecke, Alex Waibel
Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213
{yang+, wyyang, denecke, waibel}@cs.cmu.edu

Abstract

In this paper, we present our efforts towards developing an intelligent tourist system. The system is equipped with a unique combination of sensors and software. The hardware includes two computers, a GPS receiver, a lapel microphone plus an earphone, a video camera and a head-mounted display. This combination enables a multimodal interface to take advantage of speech and gesture input to provide assistance for a tourist. The software supports natural language processing, speech recognition, machine translation, handwriting recognition and multimodal fusion. A vision module is trained to locate and read written language, is able to adapt to to new environments, and is able to interpret intentions offered by the user, such as a spoken clarification or pointing gesture. We illustrate the applications of the system using two examples.

1 Introduction

A tourist faces many challenges in unfamiliar territory. Unfamiliar geography makes it difficult to navigate streets and identify landmarks. Unfamiliar language makes it difficult to read signs, take a taxi, order food, and understand the comments of passers by. Recent technological advances have made wearable computers available which could be used to ease the plight of tourists. Wearable computers can "see" as the tourist sees, "hear" as the tourist hears, and travel along with the tourist. With accessing local database and the Internet, the system might be able to have better knowledge of the environment than the tourist. This makes them excellent platforms for tourist applications. Furthermore, the mobile computing technology has made it possible for wearable computers to access information from any location.

In this paper, we present our efforts towards developing Smart Sight, an intelligent tourist assistant system. The research is to employ mobile computers to alleviate the language barrier, provide navigation assistance, and to handle queries posed and answered in natural language. The system is equipped with a unique combination of sensors and software. The tourist location is derived from a GPS (Global Positioning System) receiver. A lapel microphone plus an earphone allows for speech input and output. A video camera provides visual capabilities. This combination enables a multimodal interface to take advantage of speech and gesture input to provide assistance. For example, a tourist in a foreign land may stand in front of an information sign, circle the text and ask "what does this mean?" - for which the language translation module can then offer an informative interpretation. If there is relevant information in an online database, the information can be retrieved and be presented to the user.

This work is related to augmented reality and multimodal human computer interaction. The term augmented reality has been used to refer to enrichment of the real world with a complementary virtual world [10; 5; 4; 11]. The augmented reality systems use a see-through head-mounted display that overlays graphics and sound on a user's real vision and audition. These systems provide users with visual information that is tied to the physical, and enhance the real world by superposing additional information onto it. Multimodal signal interpretation provides a natural and flexible way for human computer interaction in a mobile environment. Multimodal interfaces consider all available human communication signals and cues rather than one alone, to better and more flexibly interpret and process human intent in communication. The multimodal signals include speech, handwriting, gesture, pointing, spelling, eye-gaze, face-location and head pose, etc.. The effectiveness of multimodal human-computer interaction has been investigated by many researchers [8; 1; 7; 9]. Over the last few years, we have focused on developing sensible and useful user interfaces to support multimodal human-computer interaction [13; 14; 16].

The remainder of this paper is organized as follows. Section 2 describes system architecture. Section 3 introduces adaptive multimodal server. Section 4 addresses problem of natural language understanding. Section 5 discusses applications of the system using two examples. Section 6 summarizes the paper.

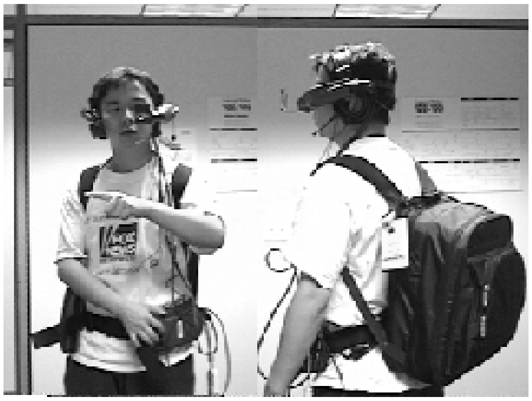


Figure 1: The system setup

2 System Architecture

A challenge for tourist applications is how to minimize system weight and bulkiness. As a suitable compromise between basic requirements of a tourist and mobility, we have employed a combination of a laptop and a wearable computer. Figure 2 shows both front view and side view of a tourist with the Smart Sight system. The system uses two computers, the Thinkpad 600 (Pentium II 333MHz) and the Xybernaut MAIV (Pentium 233Mhz), to support different tasks. The Thinkpad handles language translation task, and supports multimodal server as well map server. The Xybernaut MAIV handles the microphone input, the head mounted display, and the miniature camera input. The Thinkpad is in the backpack and the Xybernaut MAIV is on the waist belt. Two computers are connected via Ethernet using 3Com Ethernet PCMCIA cards. The hardware architecture is shown Figure 2. The reason of using two computers is partially because insufficient computing power and hardware/software support from a single computer. It is expected that we could use only one computer to accomplish the same task in the future.

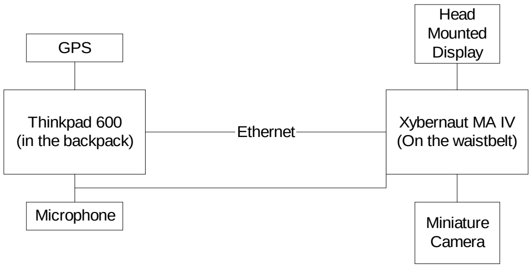

GPS
Head Mounted Display
Thinkpad 600 (in the backpack)
Ethernet
Xybernaut MA IV (On the waistbelt)
Microphone
Miniature Camera

Figure 2: The hardware architecture

Figure 3 shows the software configuration of the system. The multimodal server running on the Thinkpad controls the GPS server, the Map server, the Dialogue Server and the Speech Translation System. The following components are on the Xybernaut: the speech recognizer, the gesture recognizer, the OCR module and image processing module. The results from both the Thinkpad and Xybernaut MAIV are displayed on the head mounted display.

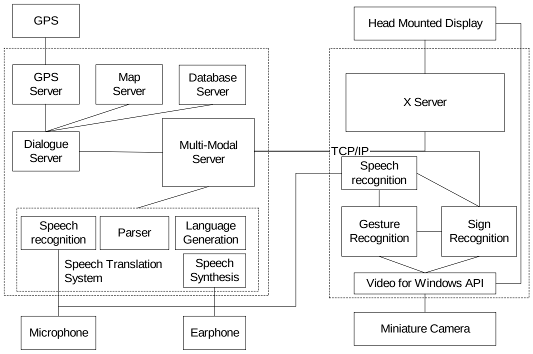


Figure 3: The software configuration

3 Adaptive Multimodal Server

The multimodal input signals are processed and interpreted by a multimodal server. The multimodal server is to provide multimodal interpretation of speech, handwriting and gestures to its clients. We have demonstrated the feasibility of the multimodal server in both web [15] and wearable [17] applications. In a wearable computer application, wearable computers can be weakly interconnected by low-speed wireless networks, or disconnected for some reasons. When the computer is connected to network, it can share a variety of resources via network. When the computer is disconnected from network, it should be able to perform it's task at least at a minimum level. It is desirable that the system has the ability to dynamically select its network service based on cost and performance requirements. This can be achieved by a dual server structure as shown in Figure 4. The multimodal inputs can be processed locally, or remotely, or partially local and partially remote.

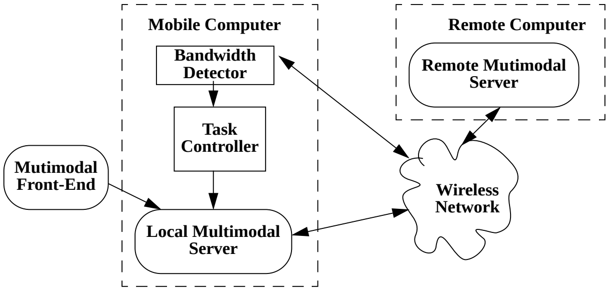


Figure 4: The dual server structure

Each server has a similar architecture as shown in Figure 5. The server can perform recognition for the modalities of speech, handwriting, and pen gesture, and interpret the recognition results. From the communication

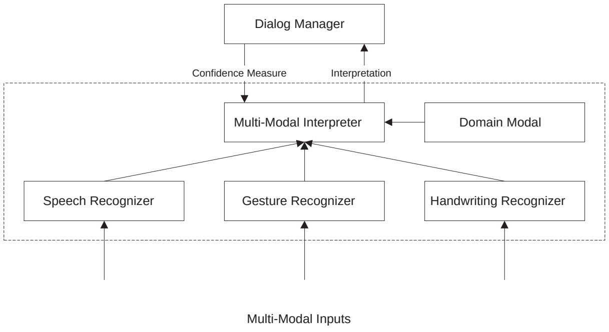


Figure 5: The multimodal server architecture

point of view, we only need to handle speech and pen inputs for modalities of speech, handwriting and gesture. In the next section we will propose a way of balancing the use of available bandwidth and computing resources.

The multimodal server is supported by the Multimodal Toolkit [12], which includes a library of software components that can be assembled to create multimodal applications. This library contains speech and pen input recorders and recognizers, multimodal event handlers, interprocess communication facilities, and a user interface in the form of a Java applet. The user interface includes ready-made objects that handle input capture and synchronization, communicate with speech and pen input recognizers, and direct the control flow of the multimodal interpretation process. This modularity permits multimodal application developers to customize each system component separately or even replace certain modules if the need arises. The distributed nature of the framework serves to spread the computational load among multiple machines, improve the responsiveness of the system, and allow resource sharing using a client/server architecture. Each major component of the system runs as a separate process which can be hosted on a different machine if necessary. The Multimodal Toolkit includes a communication layer that presents an abstract interface for interprocess communication, hiding all the details of network protocols and synchronization.

4 Natural Language Understanding Component

The system is based on a client server architecture. Currently implemented servers include a map server used for calculation and display of paths and sights, a database server using an SQL server accessing any domain-related information, and a date and time server which is also used to resolve and generate deictic time expressions such as `tomorrow` and `three hours from now`, and a GPS server.

In addition to the client/server communication scheme, there is a message passing communication mechanism that allows components to generate information. This avoids constantly polling information providing modules such as the GPS server.

4.1 The Parser

The spoken language input is analyzed by the MISO parser [6] using semantic grammars for robust speech analysis and interpretation. Although the generated parse tree contains the semantic information associated with the interpretation of the sentence, its form is not independent of the syntactic structure of the sentence. For this reason, a semantic construction algorithm converts the parse tree into a normalized partial representation, or more specifically, in a set of typed feature structure [2]. Each feature structure represents either an object in the domain (as might be expressed by a noun phrase) or a unary, binary or ternary relation between objects (as might be expressed by verbal or adjectival phrases).

4.2 The Dialogue Manager

The dialogue manager combines functionality to control and access the different servers with the original dialogue control. In addition to the server modules, the dialogue manager has control over a four layered dialogue history. The four layers contain the input representations (such as text, or gesture hypothesis), the parse trees as generated by MISO, the normalized semantic representations as generated by the semantic construction, and objects referring expressions might refer to, respectively. The dialogue history is constructed during the dialogue and may contain hierarchical structure to correctly model clarification and subdialogues. A declarative knowledge base called *type hierarchy* (figure 6) holds information on the domain in form of IS-A and IS-PART-OF relations. This knowledge base provides a typing discipline for the typed representations used in the dialogue history. The only representations used are based on typed feature structures and their generalizations.

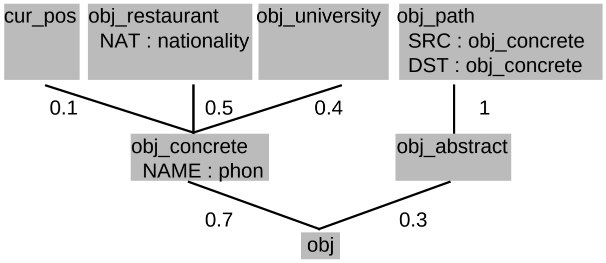


Figure 6: A part of the type hierarchy and its appropriateness conditions used in the map application. The least specific type is at the bottom of the tree.

The dialogue manager is programmable by a set of expert-system style rules. Rules may contain typed variables that range over the currently active representations in the dialogue history [3]. These rules are the only instance that defines the behavior of the system. This allows for easy adaptation to new input modalities or fast deployment of the system in new situations. In addition, since our representations are theoretically well-founded and typed, off-line type checking can be used to

$$\begin{bmatrix} \text{obj\_museum} & \left\{ \begin{bmatrix} \text{obj\_museum(2)} & \\ \text{HREF} & \text{http://www.warhol.org/warhol/warhol.html} \end{bmatrix} \right\} \\ \text{NAME} & \text{string*} \left\{ \begin{array}{l} \text{" carnegie museum of natural history "(1)} \\ \text{" andy warhol museum "(2)} \\ \text{" fort pitt museum "(3)} \end{array} \right\} \\ \text{ADDR} & \begin{bmatrix} \text{address} & & \\ \text{STR-NAME} & \text{string*} & \left\{ \begin{array}{l} \text{" forbes ave" (1,3)} \\ \text{" sandusky " (2)} \end{array} \right\} \\ \text{STR-NUMBER} & \text{int*} & \left\{ \begin{array}{l} 4400(1) \\ 117(2) \\ 100(3) \end{array} \right\} \end{bmatrix} \end{bmatrix}$$

Figure 7: Three underspecified feature structures representing the objects referred to by the NPs "the museum", "the Beehive" and "Primanti Brothers". There are two objects called "Beehive" in our data base, one being a cinema, the other one a cafe. Moreover, we find three different museums and three restaurants called "Primanti Brothers".

detect errors in the specification of the rules, a mechanism that is not available to systems such as GALAXY which are based on simpler representation formalisms such as frames, slots and fillers.

The rules are used to act as constraints to determine the relationship between simultaneously occurring input events (such as deictic references accompanied by a gesture) and sequential input events (such as answers to questions). The dialogue manager tries to relate the active discourse information with a set of predefined dialogue goals. Dialogue goals are equally expressed in typed feature structures and serve as informational lower bounds. If the available information is not sufficient to uniquely determine a dialogue goal, the dialogue manager generates discriminating information of the goals which is served as a basis for a clarification question. Additional information will then be used to disambiguate the dialogue goal. In a second step, the information that is required by the dialogue goal is sought by the dialogue system. This can be compared with a form filling approach with the important difference that over- and underspecified information also may be incorporated in the goal representation in order to allow for more flexible question and answering. An example for such an underspecified representation that serves as the basis for generating clarification questions is shown in figure 7.

Another advantage of using rules as constraints to relate different dialogue acts is the simple integration of gestures as answers in dialogue. Note that discriminating information can be generated over a set of goals as well as a set of representations of objects or actions, since in each case the representations are typed feature structures. Once the information related to the dialogue has been established, a procedure call associated with the dialogue goal is executed and the goal related information is passed on to the procedure.

The reasoning in the dialogue system is considerably fast. If no database access or other input/output operation is involved, a question can be generated as fast as half a second after receiving the input event. This time measure includes parsing, semantic construction, dialogue processing and generation. Since the input is event-based and time-stamped, no information is lost in times natural language and dialogue processing take over system control.

5 Application Examples

The Smart Sight system can help a tourist in many ways. Two applications are currently under development.

5.1 Tourist Diary

When people go vacations, they are excited about what they have seen. They take pictures to help them to remember their experience and share it with their family members or friends. However, not all the people are good at organizing their memories. Figure 8 shows a fact that among the people who take photos on vacation, 40% rarely or never put those photos into an album. It is desirable to have a system help people to organize their memories before their enthusiasm disappears.

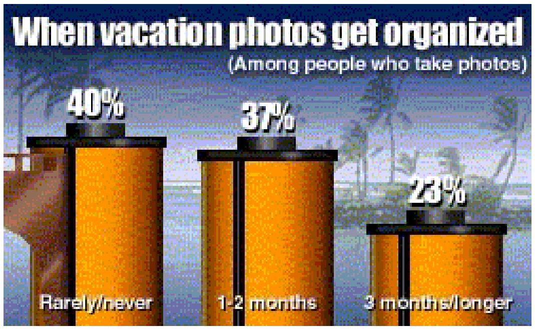


Figure 8: Disorganized memories (Opinion Research for Globus & Cosmos tour operator)

"Tourist Diary" is to help a tourist to organize his/her trip experience by multimodal interaction. The system is activated by voice commands. A tourist can request to take a picture or a video clip. He/she can add a caption to the picture and/or dictate his/her diary. Figure 9 shows the software flow chart of the system. When the tourist arrives at a point of interest, the system will log the position and time retrieved from the GPS and system clock. The tourist can ask the system to take a picture or digital video clip, and dictate the description or comments during his/her visit at the point of interest. When a local database is available, the system could also retrieve background information based on the position. After the tourist finished the site or the day, he/she could ask the system to generate an HTML document based on the stored information. He/she could then easily publish it on the web or send it to other people via email. Figure 10 shows screen shoots of the system.

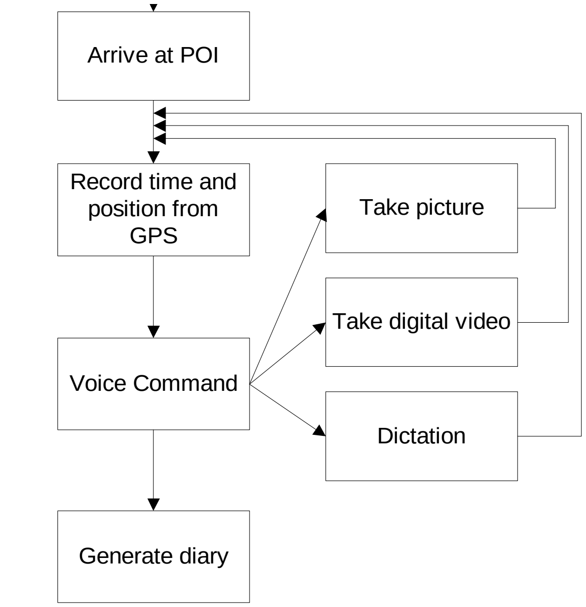


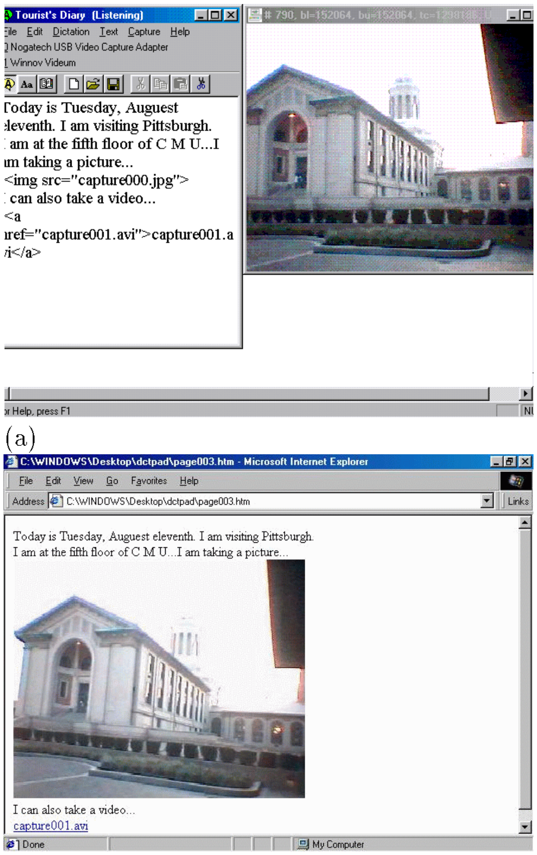
Figure 9: The software flow chart of tourist diary

5.2 "Is This an Interesting Tourist Area"

One of the most common problems in both navigation and sightseeing, typical activities of the tourist, is the identification of landmarks. How often does the average tourist point to a building and ask its name or purpose, only to have the question go unanswered? This problem could be solved by locating the tourist location and database search. If there is a sign for the landmark, the landmark can be identified by sign recognition instead of database search.

Although OCR technology has been widely used in many applications, sign identification is not a trivial problem. Challenges include difficulty in segmentation and low resolution of video image. Automatic segmentation of signs is, in principle, impossible to accomplish because they are embedded in the graphics. Fortunately, we can solve this problem by multimodal interaction in this particular application. Since a human is in the loop, the user can tell the system where he/she is interested by combination of speech and gesture. The tourist can input the sign to the system using the camera. Then he/she could ask the system "Is this an interesting tourist area?" or "Does this sign warn of a hazardous area?" with the gesture circling the location where the sign is.

The low resolution image can be enhanced by the super resolution method before the system performs the OCR. The basic idea is to take advantage of a sequence of images. The basic theory behind our approach is that of inverse graphics. That is, given a sequence of images, we want to find the ground truth (surface) that would have generated them. The most difficult part of this process is recovering the motion for each image. To do this, we will register all the images with respect to a reference image to an accuracy of a small fraction of a pixel; this registration will tell the system how an image maps onto a higher resolution image that is needed.


Tourist's Diary (Listening)
File Edit Dictation Text Capture Help
0 Nogatech USB Video Capture Adapter
1 Winnov Videum
# 790, bl=152064, bu=152064, tc=1298186
or Help, press F1
NU

Today is Tuesday, Auguest
eleventh. I am visiting Pittsburgh.
I am at the fifth floor of C M U...I
am taking a picture...
<img src="capture000.jpg">
I can also take a video...
<a
href="capture001.avi">capture001.a
vi</a>

(a)

C:\WINDOWS\Desktop\dctpad\page003.htm - Microsoft Internet Explorer
File Edit View Go Favorites Help
Address C:\WINDOWS\Desktop\dctpad\page003.htm
Links
Done
My Computer

Today is Tuesday, Auguest eleventh. I am visiting Pittsburgh.
I am at the fifth floor of C M U...I am taking a picture...

I can also take a video...
capture001.avi

(b)

Figure 10: Screen shoots of "Tourist Diary": (a) working process (b) html format result.

The system inputs the enhanced image into the OCR module and displays the result in the head mounted display. Figure 11 shows an example of recognizing a door label.

6 Conclusion

We have presented our efforts towards developing an intelligent tourist system. The system takes advantages of multimodal interaction and wireless communication. The system is equipped with a unique combination of sensors and software. The hardware includes computers, a GPS receiver, a microphone and an earphone, a video camera and a head-mounted display. The software supports natural language processing, speech recognition, machine translation, handwriting recognition and multimodal fusion. This combination enables a multimodal interface to take advantage of speech and gesture input to provide assistance for a tourist. We are currently working on system integration and developing various new applications.

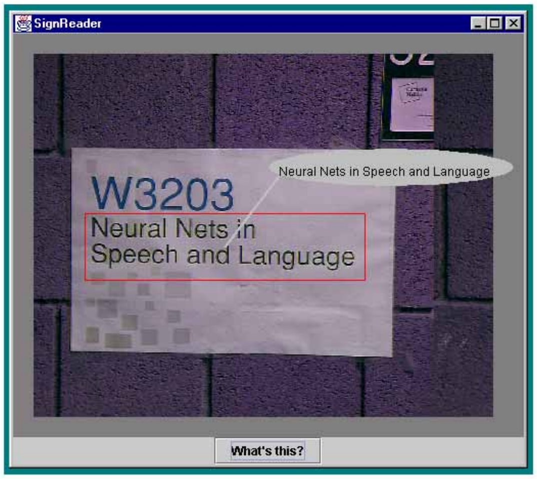


Figure 11: An example of sign recognition

Acknowledgements

We would like to thank Daniel Kiecza and Edmund Wong for their support to this project. We would also like to thank some colleagues in Interactive Systems Lab for technical support and discussions.

References

[1] Ando, H., Kitahara, Y., and Hataoka, N., ”Evaluation of multimodal interface using spoken language and pointing gesture on interior design system,” *Proc. ICSLP'94*, Vol. 2, pp. 567-570, Yokohama, Japan.

[2] Carpenter, B. *The Logic of Typed Feature Structures.* Cambridge University Press, 1992.

[3] Denecke, M., *A Programmable Multi-Blackboard Architecture for Dialogue Processing Systems.* Proceedings of the Workshop on Spoken Dialogue Processing, ACL/EACL, Madrid, Spain, 1997.

[4] Bajura, M. and Neumann, U. ”Dynamic Registration and Correction in Augmented Reality Systems.” Proc. VRAIS '95 (Virtual Reality Annual International Symp.), IEEE Computer Society Press. Los Alamitos, CA, 189-196.

[5] Caudell, T. and Mizell, D. (1992). ”Augmented Reality: An Application of Heads-Up Display Technology to Manual Manufacturing Processes.” Proc. Hawaii International Conf. on Systems Science, Vol. 2, 659-669.

[6] Marsal Gavalda and A. Waibel. *Grwoing Semantic Grammars* Proceedings of ACL/ Coling 1998, Montreal, Canada.

[7] Nakagawa, S. and Zhang, J.X., ”An input interface with speech and touch screen,” *Trans. Inst. Elec. Eng. Jpn. C (Japan)*, Vol. 114-C, No. 10, pp. 1009-1017, 1994.

[8] Nishimoto, T., Shida, N., Kobayashi, T., and Shirai, K., ”Multimodal drawing tool using speech, mouse and keyboard,” *Proc. ICSLP'94*, Vol. 3, pp. 1287-1290, Yokohama, Japan.

[9] Oviatt, S.L., Cohen, P.R., and Wang, M., ”Toward interface design for human language technology: modality and structure as determinants of linguistic complexity,” *Speech Communication (Netherlands)*, Vol. 15, Nos. 3-4, pp. 283-300, 1994.

[10] Robinett, W. ”Synthetic Experience: A Taxonomy.” Presence: Teleoperators and Virtual Environments, 1(2), Summer 1992.

[11] Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R. and Pentland, A. ”Augmented Reality Through Wearable Computing.” Presence 6(4), 1997.

[12] Vo. T.M., ”A Framework and Toolkit for the Construction of Multimodal Learning Interfaces,” Ph.D. Dissertation CMU-CS-98-129, Carnegie Mellon University (April 1998).

[13] Waibel, A., Vo, M.T., Duchnowski, P., and Manke, S., ”Multimodal Interfaces,” Artificial Intelligence Review, Special Volume on Integration of Natural Language and Vision Processing, McKevitt, P. (Ed.), Vol. 10, Nos. 3-4, 1995.

[14] Waibel, A., Suhm, B., Vo, M.T. and Yang, J., ”Multimodal Interfaces for Multimedia Information Agents,” Proceedings of 1997 ICASSP.

[15] Jing, X., Yang, J., Vo, M. and Waibel, A., ”Java Front-end for Web-based Multimodal Human-computer Interaction,” Proceedings of Workshop on Perceptual User Interfaces , pp. 78-81.

[16] Yang, J., Stiefelhagen, R., Meier, U. and Waibel, A.,”Visual Tracking for Multimodal Human Computer Interaction,” Proceedings of CHI 98, pp. 40-147.

[17] Yang, J., Holtz, W., Yang, W. and Vo, M., ”An Adaptive Multimodal Interface for Wireless Applications,” Proceedings of International Symposium on Wearable Computers , Pittsburgh, PA, Oct. 19-20, 1998.