MULTIMODAL INTERFACES FOR MULTIMEDIA INFORMATION AGENTS

Alex Waibel, Bernhard Suhm, Minh Tue Vo and Jie Yang

Interactive Systems Laboratories
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15217 (USA)
{waibel,bsuhm,tue,yang+}@cs.cmu.edu

ABSTRACT

When humans communicate they take advantage of a rich spectrum of cues. Some are verbal and acoustic. Some are non-verbal and non-acoustic. Signal processing technology has devoted much attention to the recognition of speech, as a single human communication signal. Most other complementary communication cues, however, remain unexplored and unused in human-computer interaction. In this paper we show that the addition of non-acoustic or non-verbal cues can significantly enhance robustness, flexibility, naturalness and performance of human-computer interaction. We demonstrate computer agents that use speech, gesture, handwriting, pointing, spelling *jointly* for more robust, natural and flexible human-computer interaction in the various tasks of an information worker: information creation, access, manipulation or dissemination.

1. INTRODUCTION

Human-computer interfaces today are limited and inflexible and do not take advantage of the many communication channels humans use to communicate verbal and non-verbal ways. Humans speak, point, gesture, write, fixate, use facial expressions, head-motion, eye-contact, etc. to express ideas, intentions and feelings. To build computer systems that operate more flexibly and robustly, and that are more intuitive to use by anyone, computer interfaces must be able to understand and process this multiplicity of human communication cues. Recently many research projects have been conducted to address this problem (e.g. [1], [2], [3], [4]). The need for more intuitive interfaces becomes all the more pressing as multimedia presentation, that is output, is becoming commonly available.

In this paper we describe information agents that use multimodal interfaces to access, manipulate, create and disseminate information. In particular, we seek effective methods by which human users can easily interact with a computer by speaking, pointing, drawing, spelling, etc. We show how cross-modal cues can be used effectively to recover from errors or miscommunications, since confusions in one modality usually are unambiguous in another. We describe how multiple modalities can cooperate and complement each other in various phases of the preparation of multimedia documents. Moreover, the added flexibility and robustness significantly increases reliability and naturalness, and with it the usability and acceptance of information systems. We present QuickDoc as a prototypical application that combines our multimodal subsystems in a simple, but powerful way.

2. INTERPRETATION OF MULTIMODAL INPUT

2.1. Multimodal Components

Our labs have implemented recognizers and processing modules for various input modalities, most notably speech, pen input, and face tracking. These modality processors constitute the basic components of all our multimodal systems.

Speech

Our speech recognition subsystem is based on the recognition front-end of the JANUS speech translation system [5] which is capable of processing speaker-independent, spontaneous speech. The recognizer can be adapted to any task domain by retraining the language models and possibly tuning the acoustic models if the domain involves special vocabulary. We also have a high-performance, real-time continuous spelling recognizer for large lists of 100,000 or more names.

Gesture

Our approach to pen-based gesture recognition is to decompose pen strokes into sequences of basic shapes such as line, arc, arrow, circle, cross...[6] The same gesture shape may mean different things depending on the surrounding context, hence each gesture component is augmented by *gesture contexts* indicating spatial relationships between the gesture and nearby objects in the user interface.

Handwriting

Our MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting at a recognition rate of 94% on a 20,000-word vocabulary. We employ simple heuristics to decide when to invoke handwriting recognition on pen input, e.g., when the gesture recognizer cannot identify the input strokes as basic shapes.

Face Tracking

We base our face-tracking subsystem on face color clustering and motion detection [8]. The face tracker can control a pan-tilt-zoom camera to follow a freely moving person, producing a constant-sized image of the face area in real time. Another software layer built on top of this face tracker can identify the eyes and other face features, build a model of the face, and estimate the gaze direction to track the head pose [9].

2.2. Joint Interpretation

In order to make sense of input from all available sources, we need a multimodal interpreter capable of producing an interpreta-

0-8186-7919-0/97 $10.00 © 1997 IEEE

tion of user intent (e.g., a command to execute in the application interface) from the output of the modality processors.

In our joint interpretation scheme, the user intent is represented by a frame consisting of slots specifying pieces of information such as the action to carry out or the parameters for that action. Recognition output from the modality processors are parsed into partially filled frames that are merged together to produce the combined interpretation as described in [6]. This technique leads to uniform handling of high-level information from all input sources, which is very important for modularity and extensibility. To add another input modality we need only provide a module to convert low-level recognizer output to a partially filled frame to be merged with others. In addition, context information can be retained across input events by merging with previous interpretation frames.

3. MULTIMODAL ERROR CORRECTION

Intensive research over recent years has boosted the performance of speech recognition technology significantly. Nevertheless it is widely believed that recognition performance will remain limited, at least for the foreseeable future. The potential for error with any speech interface will be balanced by the redundant and alternate ability to enter information or queries by spelled, handwritten or typed input. Rather than fighting a particular recognition error the user can therefore quickly circumvent the problem by choosing a different modality. Such alternates not only eliminate user frustration with potentially recurring errors, but also provide an avenue for background learning and adapting.

3.1. Multimodal Interactive Error Repair

Our approach is to have the user collaborate with the system in recovering from interpretation errors [10]. First, errors have to be identified, either initiated by the system, for instance based on some confidence measure, or by the user. If a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.

The rationale for this multimodal approach to error recovery is twofold. First, it exploits the fact that different input modalities are orthogonal: words which are confusable in one modality, can be disambiguated in a different modality (e.g. "road" and "rote" spoken versus spelled). Secondly, recent studies [11] show that switching modality after system misinterpretation alleviates user frustration.

3.2. Error Repair for Multimedia Information Agents

The design of speech user interfaces is constrained by the context of the application, and so is the design of error repair. The application context varies along several dimensions: available modalities (speech only versus GUI), what is repaired (isolated words, phrases, sentences), goal of repair is (get verbatim every word versus get the intended action), dialogue metaphor (command controls versus conversational). For multimedia information agents, we see the following two application contexts as particularly relevant:

- (Mixed initiative) Spoken dialogue, for example in user queries and command control of the application. In this context, a GUI may be available, repair will typically be performed at the level of phrases, the goal of repair is to initiate the intended action (semantic repair), and a conversational dialogue is desirable.
- Dictation, for example to entry in fields, or to dictate reports. In this context, a GUI is very appropriate, repair will occur from isolated word to sentence level, the goal of repair is to get each word correct (verbatim repair).

We have built prototypical speech user interfaces with error repair capabilities for dictation and form filling tasks. In this application context, requiring the user to identify errors by highlighting them is appropriate, since the user naturally focuses on the actual recognition hypothesis. For error correction, we offer to repeat the input (potentially switching to another modality), and pen gestures to indicate where to insert words, or which word to delete. In form filling tasks, knowledge of the current field can provide powerful constraints for recognition by restricting vocabulary and language model accordingly.

3.3. Evaluating Interactive Error Repair

Given a set of error recovery methods feasible with current technology in a given application context, a crucial issue is to predict which methods a user will prefer. Based on the assumption of a rational user [13], we use the time to complete some input, including the time needed to correct errors, as an objective and easily quantifiable measure for the effectiveness of different error recovery methods, and main predictor of user preference. By assuming stochastic independence of the various repair attempts, the expected accuracy after a certain number of repair attempts can be estimated as a geometric series. Given a high level of accuracy sufficient for the application (e.g. 99%), the number of repair attempts can be estimated, and thus the total time required to input including repair [12].

A pilot evaluation on a form filling task [12] suggests that given current technology, repair by spelling and handwriting can be very effective, and significantly better than standard choice from N-best alternatives. Using the context of a repair, e.g. in rescoring N-best lists obtained by decoding repair input, can substantially improve the accuracy of repair. Although results are still preliminary, they show that our multimodal approach to interactive error recovery is very promising.

4. MULTIMODAL INFORMATION AGENTS

The objective for multmodal information agents is to quickly access, create, manipulate, and disseminate multimedia information. A multimodal agent can label multimedia information with appropriate classifications by voice or/and gesture annotations. All annotations are incorporated with original information. The agent will automatically attach hotlinks on the world-wide-web to broaden information sources. The agent then will generate an HTML format report. The agent can search all the information in the report including voice and gesture annotations and voice mails. The agent can repair errors and add more information. Finally, the agent can disseminate the report in electronic form or printed form.

4.1. Information Access

Multimodal information agents can speed up navigation of Information space and queries of a database. A city map, for example, can be queried by a combination of gesture, handwriting and speech. A query such as: "Show me all hospitals in this <circling-gesture> area," or "How far is from here to there <tracing-gesture>?" will access information that would otherwise have to be assembled and combined painstakingly in a series of steps. Multimodal agents will recognize gesture and speech concepts and map them onto a semantic representation. The semantic representation can be filled interchangeably by both gestures and verbal phrases using pattern recognition and chart parsing techniques. The parsing result is then sent to query the database. If interpretation errors occur, the user corrects them using multimodal correction techniques. The whole process is transparent to the user.

Searching multimedia data can be viewed as natural extension of multimodal database queries. For instance, voice annotations or video clips added during user interaction can be interpreted and indexed using multimodal interpretation components, and then accessed using database queries techniques.

4.2. Information Creation

A user often needs to add information to a database or modify existed information. Multimodal agents are also useful in such a process. For example, a doctor needs to mark his/her diagnoses on X-ray, CT, MR images and make comments on the case. He/she can use speech, gesture, and handwriting to do this. Similarly iconic gestures may be used on a touch sensitive display or writing pad, to quickly enter resources or objects in a database, such as a map, in non-verbal ways. Iconic gestures can be defined to represent objects that would be entered in the database. In the same way, handwriting symbols can be used to mark or tag objects on a screen. Multimodal input also need not always be interpreted, but can often be stored for later review. Thus a voice annotation, or a circling gesture can be simply stored to go along with an image for later review and retrieval. The efficiency of multimodal agents for information creation can illustrated by the following example. The scenario is an image analyst sitting in the front of a computer and classifying various targets. Images can be popped up quickly and in rapid succession and the analyst provides a quick classification. The image and its classification are entered into a database along with the analyst's spoken and gestured annotation. With the human in the loop to make decisions on intelligence relevance, but with the availability of a better (multimodal) interface, to do so quickly, the job of data creation and manipulation can be carried out with greater reliability and efficiency compared to a fully automated system.

4.3. Information Manipulation

Information will not only be requested to viewed, but will also need to be updated. A multimodal multimedia information system can offer new ways to maintain the underlying databases: the user can use speech, gesture, writing, spelling to update database information efficiently. manipulating visual representations on the display rather than having to learn and use a separate database manipulation system. Gestures play an important role in specifying object parameters (e.g. different iconic shapes to represent different object types) and spatial constraints (e.g., location and extent of objects), while speech is useful in specifying parameters not easily expressed visually. For example he/she might circle a bridge on a map and say "this bridge has been destroyed", or might cross out an object without words, etc.

In some situations, the answer for a query cannot be directly obtained from the database and the information from the database has to be future processed. For example, the query "Show me the nearest restaurant" requires not only database access but also other computations.

4.4. Information Dissemination

In addition to information access, creation and manipulation, the product has to be packaged and readied for later viewing. The world wide web can serve as a natural medium to disseminate multimedia information. The multimodal input signals are interpreted and compiled, and an HTML format report is automatically generated on the fly. It is then automatically available for viewing together with the original aligned speech waveform attached to it as hotlinks. Multimodal interaction can also be used to manipulate objects in a report, such as positioning objects on a page appropriately, or attaching appropriate hotlinks.

To quickly generate a multimedia report, speech dictation can be combined with point and click actions to arrange reports on the fly. Such a multimodal report generation would be faster than typing and can generate a hierarchy of facts and notions. The generation thus would consist of some dictation or typing, some multimodal object manipulation ("move this <point> here <point> and link it as an explanation for that <click>") and some handwriting or drawings that are to be attached. The result of these communicative acts will compile into HTML code, that is ready to be transferred. Time savings are possible in several ways: the report generation can be done more efficiently by speech, gesture, or handwriting, and by using references to already existing multimedia stickies, factoids, video footage or explanations, and the recipient of the report can probe and browse a hierarchical multimedia report in a non-linear fashion more efficiently, calling up aspects of the assessment as needed.

4.5. Controlling the Interface

Although the goal of a multimodal interface is to provide as transparent an access to functionality as possible, interaction to control various aspects of the system can't be eliminated.

For example for navigation within visually presented image data, the user typically modifies the view by operations such as zooming and panning. The integration of speech and gesture increases flexibility by allowing different attributes of an operation to be specified in whatever modality is more appropriate for each attribute; for instance, the effect of a spoken "zoom in to this <circle> area" accompanied by a circle drawn around the desired area is much more difficult to achieve using a single input modality

Additionally, multiple modalities can be used to enhance the reliability. For example, a gaze tracker can be used to detect the user's focus of attention. When the user is looking at a window on a screen, the window will be highlighted. No action is taken unless the user uses a voice command to confirm the selection. The voice commands could be ``select this window" or ``close window", etc. This can reduce unintended actions caused by the user randomly looking around.

5. THE QUICKDOC APPLICATION

QuickDoc is an example of an application that combines our multimodal subsystems in a simple but powerful way, letting the user perform a repetitive task with speed and convenience. The task is for a doctor to go through a series of images such as X-rays or computer-aided tomography scans, quickly identify an anomalous area, label the area with the name of a disease or condition, and attach relevant comments. The end product is an HTML report that summarizes the doctor's findings in a compact table listing the annotated images, the corresponding preliminary diagnoses, and automatically generated hotlinks to relevant sites based on the diagnoses.

The QuickDoc user labels each presented image by circling an area on the image (drawing directly on a touch-sensitive screen) and speaking a disease name as well as a percentage representing the confidence level of the diagnosis. The gesture recognizer identifies the circle and generates an area marker on the image, attaching other parts of the gesture as written annotations (a future version may run the handwriting recognizer to turn those into text). The user can issue a spoken command to initiate a voice annotation attachment; the recorded audio file is also run through the speech recognizer to produce an automatic transcription. When all the images have been processed, another spoken command causes the HTML report to be generated. Optionally the collected data can also be edited in a multimodal dictation and repair facility before being processed into HTML.

The generated report contains thumbnails of the annotated images accompanied by disease label, confidence level, location in the image (automatically extracted from the circling gesture), and voice annotation playback icon if applicable. The thumbnail is linked to a more detailed page containing the full-size image and the automatic voice transcription. The name of the disease is looked up in a database and turned into a hotlink to a page listing Web sites relevant to that disease. A keyword search mode allows to search spoken keywords in the text of the report and the automatic voice transcriptions; entries containing hits are then highlighted in the report.

QuickDoc thus embodies multimodal creation, manipulation, dissemination, and access of multimedia information in a simple yet surprisingly useful application.

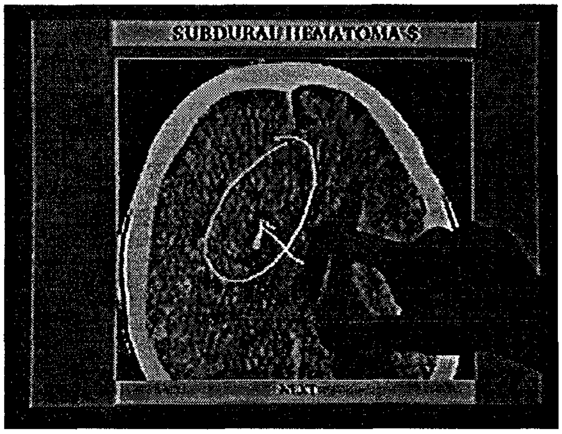

FIGURE 1: Multimodal information creation in QuickDoc: labeling some area by a circle gesture and speech ("subdural hematoma")

6. CONCLUSIONS

In this paper we have described systems that combine speech, gesture, handwriting, pointing, spelling, and other communication modalities into interfaces that are robust, flexible, and intuitive to use. We described how multimodal error recovery can ease the problem of unreliable automatic interpretation of communication modalities, in particular speech, thus removing a major obstacle in making multimodal interfaces truly usable. We demonstrated how these concepts can be combined for a more user friendly multimedia document production technology.

7. ACKNOWLEDGEMENTS

This research was sponsored by the DARPA under the Department of the Navy, Naval Research Office under grant number N00014-93-1-0806. Views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Navy or the U.S. Government.

8. REFERENCES

[1] H. Ando, Y. Kitahara and N. Hataoka: "Evaluation of Multimodal Interface Using Spoken Language and Pointing Gesture on Interior Design System", *Proc. ICSLP*, 1994, Vol. 2, pp. 567-570.

[2] T. Nishimoto, N. Shida, T. Kobayashi, K. Shirai: "Multimodal Drawing Tool Using Speech, Mouse and Keyboard", *Proc. ICSLP*, 1994, Vol. 3, pp. 1287-1290.

[3] J. Wang: "Integration of Eye-gaze, Voice and Manual Response in Multimodal User Interface", *Proc. ICSMC*, 1995, Vol. 5, pp. 3938-42.

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

[5] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L.J. Mayfield, A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel: "JANUS: Towards Multilingual Spoken Language Translation," *Proc. ARPA SLT Workshop 95* (Austin, Texas).

[6] M.T. Vo and C. Wood: "Building and application framework for speech and pen input integration in multimodal learning interfaces," *Proc. ICASSP'96* (Atlanta, GA).

[7] S. Manke, M. Finke and A. Waibel: "The Use of Dynamic Writing Information in a Connectionist On-Line Cursive Handwriting Recognition System," *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, 1994.

[8] J. Yang and A. Waibel: "A real-time face tracker," *Proc. WACV'96*, pp. 142-147.

[9] R. Stiefelhagen, J. Yang and A. Waibel: "A Modelbased Gaze Tracking System," *Proc. IEEE International Joint Symposia on Intelligence and Systems - Image, Speech & Natural Language Systems*, pp. 304-310, 1996.

[10] A.E. McNair and A. Waibel: "Improving Recognizer Acceptance through Robust, Natural Speech Repair", *Proc. ICSLP*, 1994, Vol. 3, pp. 1299-1302.

[11] S.L. Oviatt and R. VanGent: "Error Resolution during Multimodal Human-Computer Interaction", *Proc. ICSLP*, 1996.

[12] B. Suhm, B. Myers and A. Waibel: "Interactive Recovery from Speech Recognition Errors in Speech User Interfaces", *Proceedings of the International Conference on Spoken Lanuage Processing - ICSLP*, 1996.

[13] A. Newell, S. Card and P. Moran: "The Psychology of Human Computer Interaction", Lawrence Earlbaum Associates, 1983.