Multimodal Interpretation Gesture Recognition System: A Review

S.A.Chhabria
Assistant Professor,
Department of Information Technology G.H.Raisoni College Of Engineering

R.V.Dharaskar
Director, Matoshri Pratishthan's Group of Institutions, Nanded ,India

V.M.Thakre
Head Of Department of CSE
College Of Engineering, Amravati

ABSTRACT

Multimodal systems allow humans to interact with machines through multiple modalities such as speech, gesture and gaze.

Different multimodal systems that have been developed so far will be discussed in this paper. These include put that there,Cubricon,Xtra, Quickset,RIA with MIND The growing interest in multimodal interface design is inspired in large part by the goals of supporting more transparent, flexible, efficient, and powerfully expressive means of human–computer interaction than in the past. Multimodal interfaces are expected to support a wider range of diverse applications, be usable by a broader spectrum of the average population, and function more reliably under realistic and challenging usage conditions. We also describe a diverse collection of state-of-the-art multimodal systems that process users' spoken and gestural input. These applications range from map-based and virtual reality systems for engaging in simulations and training, to field medic systems for mobile use in noisy environments, to web-based transactions and standard text-editing applications that will reshape daily computing and have a significant commercial impact. To realize successful multimodal systems of the future, many key research challenges remain to be addressed. Among these challenges are the development of cognitive theories to guide multimodal system design, and the development of effective natural language processing, dialogue processing, and error-handling techniques. In addition, new multimodal systems will be needed that can function more robustly and adaptively, and with support for collaborative multiperson use.

Gesture interpretation can be seen as a way for computers to begin to understand human body language, thus building a richer bridge between machines and humans than primitive text user interfaces or even GUIs, which still limit the majority of input to keyboard and mouse. It has also become increasingly evident that the difficulties encountered in the analysis and interpretation of individual sensing modalities may be overcome by integrating them into a multimodal human–computer interface. This research can benefit from many disparate fields of study that increase our understanding of the different human communication modalities and their potential role in Human Computer Interface which can be used for handicapped persons to control their wheel-chair, expert to have computer assisted surgery, mining etc .

Keywords

Human Computer Interaction, Gesture Interpretation, Multimodality, Noise

1. INTRODUCTION

Multimodal interfaces process two or more combined user input modes in a coordinated manner with multimedia system output. Such combinations work to facilitate the overall human computer interaction experience. There is a growing interest in the design and implementation of multimodal interfaces fueled by the many inherent advantages they provide. Multimodal systems are flexible in their ability to provide users with a choice of input. They offer greater accessibility to a broad range of users. Their adaptability is apparent in their ability to switch input modes as necessary. The simultaneous input possibilities afforded by multimodal interfaces allow for more efficient input. Multimodal systems can also take advantage of mutual disambiguation to facilitate error avoidance and recovery.

Technologies used in multimodal interfaces include conventional direct-manipulation devices like the keyboard, mouse, pen and touch screen, as well as progressively more advanced recognition technologies such as speech recognition, 2D and 3D gesture recognition, and lip movement and gaze tracking. The most mature research to date integrates speech with pen, or speech with lip movement tracking. Usability studies, exploring and evaluating the human factors involved in multimodal input, provide useful insight and guidance toward the design and implementation of multimodal interfaces.

The primary goal in the design of any user interface is to facilitate the interaction between user and machine. This user-centered goal is the guiding force behind choices made in the design process. There are, of course, many system engineering issues that influence interface design decisions such as the limits of technology, schedules, proper functionality, reliability, etc.

2. EARLY MULTIMODAL INTERFACES

One of earliest multimodal interfaces illustrating the use of voice and gesture based input is Richard Bolt's "Put That There" system [BOLT80]. Subsequent multimodal interfaces of the late 1980's and early 1990's explored the use of speech input combined with conventional keyboard and mouse input. The design of these interfaces was based upon a strategy of simply adding speech to traditional graphical user interfaces (GUIs). The primary motivation for this addition of speech was a belief that the use of speech gives the user greater expressive capability, especially when interacting with visual objects and extracting information. Examples of such types of interfaces include CUBRICON [NEAL90], XTRA [WAHLSTER91].

2.1 Put-That-There

In Bolt's "Put-That-There" system, speech recognition is used in parallel with gesture recognition. [1] User interaction takes place in a media room about the size of a personal office. Visual focus is directed at a large screen display on one wall of the room. Gesture-based input is primarily the recognition of deictic arm movements in the forms of pointing at objects displayed on the screen and sweeping motions of the arm whilst pointing. In general, deictic gestures are gestures that contribute to the identification of an object (or a group of objects) by specifying their location. The gesture recognition technology used involves a space position and orientation sensing technology based on magnetic fields [BOLT80]. Speech recognition in the “Put That There” system allows for simple English sentence structures using a limited vocabulary.

The driving example scenario Bolt uses to illustrate his “Put That There” system consists of input requests for creating, customizing, copying, moving, and deleting basic geometric objects on a large screen display. Sample input speech includes the command, “Create a blue square there”. The difficulty of interpreting this request is the presence of the pronoun, *there*. In a purely speech based interface the specification of a location must be included as part of the speech command. For example, after uttering, “Create a blue square…”, location information has to follow, in the form of, “…in the center of the display.”, or perhaps, “…next to the green circle.”, (assuming referable objects, such as a green circle exist). Bolt’s system addresses this challenge by having pronouns refer to temporal arm pointing and motion gestures. This disambiguation is representative of how multimodal interfaces can cooperatively use one modality in parallel with another.

The speech utterances recognized by Bolt’s “Put-That-There” system are limited to its set of command words. This contrast with later trends which tend to build more upon natural language processing. A limited speech recognition vocabulary can be useful because it improves recognition efficiency and accuracy. Bolt’s system provides an initial step in establishing multimodal interfaces as a more natural form of human-computer interaction. This is especially evident in the user’s ability to use pronouns as they would in daily conversation, and the natural manner of pointing to an object to establish it as the subject of current discourse.

2.2 CUBRICON

An interface combining spoken and typed natural language with deictic gesture for the purposes of both input and output was designed for CUBRICON [NEAL90], a military situation assessment tool. Similar to the “Put-That-There” system, the CUBRICON interface utilizes pointing gestures to clarify references to entities based upon simultaneous natural language input. It also introduces the concept of composing and generating a multimodal language based on a dynamic knowledge base. This knowledge base is initialized and built upon via models of the user and the ongoing interaction. These dynamic models influence the generated responses and affect the display results which consist of combinations of language, maps, and graphics.[1]

In the CUBRICON architecture, natural language input is acquired via speech recognition and keyboard input.

Location coordinates are specified via a conventional mouse pointing device. An input coordinator processes these multiple input streams and combines them into a single stream which is passed on to the multimedia parser and interpreter. Building upon information from the system’s knowledge sources, the parser interprets the compound stream and passed the result on to the executor/communicator.

The CUBRICON system’s knowledge sources are comprised of:

 Lexicon *Grammar* - defines the multimodal language

 Discourse Model - dynamically maintains knowledge pertinent to the current dialog.

 User Model - aids in interpretation based on user goals and plans

 Knowledge Base - contains information related to the task space

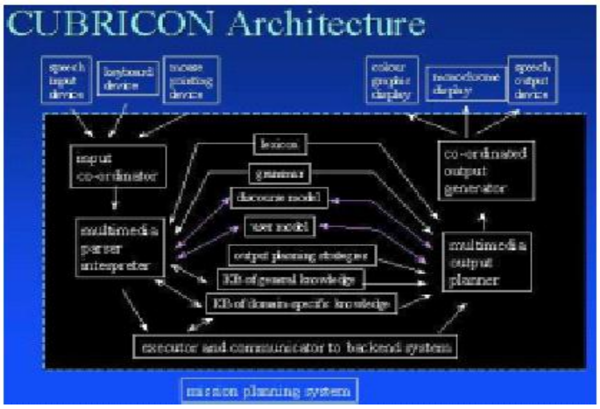


Figure 1 : CUBRICON Architecture.

CUBRICON’s multimodal language incorporates mouse input to select on-screen content, such as windows, table components, icons, and points, and spoken or written natural language, to specify an action(s) that refers back to selected objects. It builds upon “Put-That-There”, by allowing a number of point gestures in a single phrase and the combination of multiple multimodal phrase into one sentences. For instance, CUBRICON allows one to use a phrase like “Where are these items?”, while sequentially pointing to multiple elements.

The combination of speech and gesture in this manner improves the usability of either input method alone, as the two can work cooperatively to achieve greater accuracy in determining the user’s intent. Thus the interpretation of an ambiguous utterance can take advantage of the fact that only a limited set of applicable actions exist for the referenced object. Conversely, an ambiguous pointing gesture can be resolved if simultaneous natural language input reduces the applicable on-screen objects.

CUBRICON’s output is also multimodal as it integrates gesture with speech. For instance if an output refers to an icon object the icon referenced is pointed to and corresponding natural language is generated. If the object is part of an icon the containing icon is pointed to instead. If the output refers to an object that appears in multiple windows the object is weakly highlighted in each window, except for the top or selected window, in which case the icon blinks.

2.3 XTRA: An Intelligent Multimodal Interface to Expert Systems

XTRA (expert Translator) is an intelligent multimodal interface that combines natural language, graphics, and pointing for input and output. [WAHLSTER91]. Based upon a focusing gesture analysis methodology, the XTRA project constrains referents in speech to possibilities from a gesture based region. Doing so aids the system in interpretation of subsequent definite noun phrases which refer to objects located in the focused area.

An illustrative application discussed by Wahlster involves the use of XTRA to facilitate filling in a tax form. As shown in Figure 2, gesture based input and output for this application occurs in the left panel which displays pages of a tax form.

Natural language input and system response text are displayed in a panel to the right. Note that the tax form display panel is shared for both gesture based input and output.

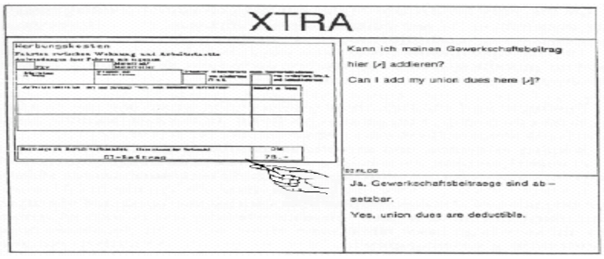
Using a mouse or similar pointing device the user can specify locations on, and areas of, the tax form. Fields that exist on a tax form page may overlap or be contained within another. Also, analogous to human-human interaction, but unlike conventional human-computer interaction, gestured-to locations are not confirmed graphically.

The granularity and interpretation of mouse-specified locations and areas depends upon the current pointing mode selected by the user. These modes are designed to simulate various types of deictic gestures commonly used in human-human conversation as follows:

Exact pointing with a pencil, Standard pointing with the index finger, Vague pointing with the entire hand, Encircling regions with an ‘@’-sign.

In addition, three types of movement gestures are considered: point, underline, and encircle. Selecting in pencil mode is similar to mouse selection in conventional WIMP-based interfaces, however, as the pointing area mode becomes less granular, mouse selections are no longer considered to occur in discrete fields.

Instead, a plausibility value is computed for each subset of the superset generated with all of the fields contained in the pointing-mode based mouse selection region. Thus a selection of multiple tax form fields as a referent could be accomplished by using the entire hand mode and using plurality in the natural language discourse.



Kann ich meinen Gewerkschaftsbeitrag hier [↗] addieren?

Can I add my union dues here [↗]?

Ja, Gewerkschaftsbeitraege sind ab – setzbar.

Yes, union dues are deductible.

Figure 2: XTRA Tax Form

Also considered by XTRA are the effects of *dialog focus*, which allows the user to sequentially or simultaneously specify a region to be the one containing another location or area.

XTRA is a foundational illustration of how dynamic user models and dialog discourse models should affect the multimodal output of a cooperative natural language and gesture based interface and vise versa.

In addition, it introduces the use of deictic gesture granularity to parallel natural gestures usage in human-human interaction. XTRA also showed a use of sequential or simultaneous pointing gestures in which one gesture establishes an area of *attention* to reduce or remove ambiguity in another gesture.

3. RECENT SPEECH BASED MULTIMODAL INTERFACES

Recent multimodal interface trends have moved away from combining speech with simple mouse and touchpad pointing, and toward the use of speech in parallel with more expressive input methods and technologies [OVIATT02]. Such recent interfaces are more powerful in their ability to utilize two recognition based inputs. Currently the most mature research in multimodal interfaces, combining two recognition based inputs, has focused on speech and pen or speech and lip recognition. For both cases keyboard and mouse input tends to not be used.

3.1 Quickset

Research into speech and pen based multimodal input began in the early 1990’s. The Quickset system, prototyped in 1994, is one of the earliest speech and pen multimodal interfaces [OVIATT02]. Quickset is a collaborative multimodal system designed to run on multiple platforms from handheld PCs to wall-sized display interfaces. In addition to integrating multiple interface components, the Quickset system is designed to work with a collection of distributed applications [COHEN97]. A Java-based implementation of Quickset was developed for the World Wide Web. The system also introduces a unification-based mechanism to analyze the meaning of multiple input mode fragments.[1]

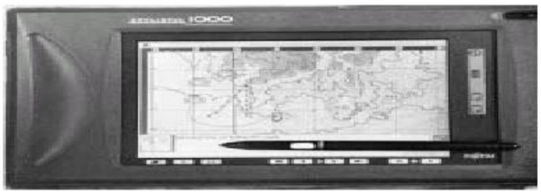


Figure 3: Quickset Handheld PC Interface

This mechanism selects the optimal joint interpretation of sequential or simultaneous input fragments. Like the CUBRICON and XTRA systems, Quickset utilizes multimodal discourse to aid in accurate interpretation of speech and gesture input. Quickset is designed as a general architecture for providing speech and pen multimodal interfaces for map-based, otherwise self contained, back-end applications [COHEN97]. The map interface provided by QuickSet displays the terrain for a specified region along with entities whose physical position lies within the region. Normal map interface capabilities such as zoom and pan are also provided. Multimodal pen and speech input allows the user to annotate the map using points, lines, and areas. The user can also use symbolic gestures to create new entities on the map while simultaneously using speech input to describe and name them. To handle the situation where background conversation or speech is not intended for the interface, Quickset only activates its speech recognition engine when the pen touches the display. The commercial speech engines used by Quickset to implement speech recognition are IBM’s Voice Type, a predecessor to the current IBM Via Voice series, and Microsoft’s Whisper engine. The pen-based gesture recognizer was written as part of the Quickset implementation and consists of a neural network and a set of hidden Markov models. The gesture recognizer recognizes a number of pen gestures including military map symbols, editing gestures, paths, areas, and taps. Quickset also provides distributed system support, speech recognition customization parameters, and multi-user collaboration.[1]

To conclude, a general framework such as Quickset facilitates multimodal interface design research by providing a flexible testing environment, in which multimodal interfaces can be developed and refined using a rapid implementation and test cycle. This environment allows researchers to acquire a better understanding of which interface modes and combinations work best with particular application paradigms.

3.2 REA & MIND

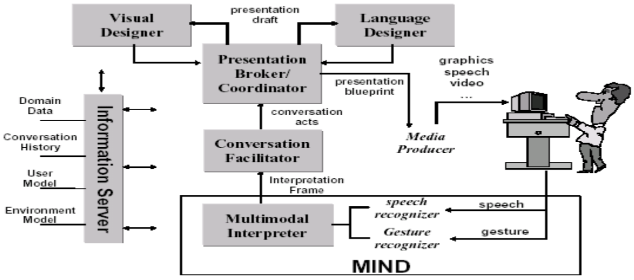


Figure 4: RIA infrastructure

Figure 4 shows RIA's main components. A user can interact with RIA using multiple input channels, such as speech and gesture. [2] To understand a user input, the multimodal interpreter exploits various contexts (e.g., conversation history) to produce an interpretation frame that captures the meanings of the input. Based on the interpretation frame, the conversation facilitator decides how RIA should act by generating a set of conversation acts (e.g., Describe information to the user). Upon receiving the conversation acts, the presentation broker sketches a presentation draft that expresses the outline of a multimedia presentation. Based on this draft, the language and visual designers work together to author a multimedia blueprint which contains the details of a fully coordinated multimedia presentation. The blueprint is then sent to the media producer to be realized. To support all components described above, an information server supplies various contextual information, including domain data (e.g., houses and cities for a real-estate application), a conversation history (e.g., detailed conversation exchanges between RIA and a user), a user model (e.g., user profiles), and an environment model (e.g., device capabilities).[2]

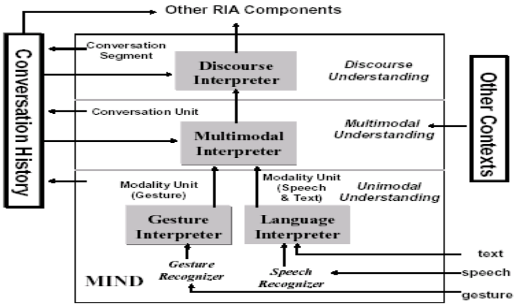


Figure 5: MIND Overview

To interpret multimodal user inputs, MIND takes three major steps as shown in Figure 5: unimodal understanding, multimodal understanding, and discourse understanding. During unimodal understanding, MIND applies modality specific recognition and understanding components (e.g., a speech recognizer and a language interpreter) to identify meanings of each unimodal input.[2][9]

During multimodal understanding, MIND combines semantic meanings of unimodal inputs and uses contexts (e.g., conversation context and domain context) to form an overall understanding of multimodal user inputs. Furthermore, MIND also identifies how an input relates to the overall conversation discourse through discourse understanding.

4. CONCLUSION

The design and implementation of multimodal interfaces is an exciting area of research in the field of human-computer interaction. Initial research in this area includes multimodal systems such as Bolt's "Put-That-There" system which combines speech and gesture, allowing users to identify and act upon referents in speech by physically pointing at their visible representations. Other early systems in this genre include the CUBRICON system, which studies the benefits of maintaining dynamic user and discourse models, to improve interpretation of gesture-based and natural language speech multimodal input, and the XTRA system that also includes the use of user and discourse models while exploring the use of variable granularity in deictic gestures involved in a point-and-speak interface model.

More recent systems include: QuickSet, a reusable map-based speech and pen multimodal interface framework that allows more complex symbol gestures for creating objects as well as spatial and pointing gestures, Human factors that need to be considered in the implementation of multimodal speech based interfaces include individual voice quality, short term memory, dialog usage structure, conversational technique, vocabulary, and speech prosody. The human factor of emotion has been the subject of recent studies that explore methods of detecting and addressing emotion during speech input analysis, and designing interfaces that avoid soliciting negative emotions.

Research and implementation of multimodal systems is fueled by the many inherent advantages they provide. Multimodal systems are flexible in their ability to provide users with choice of input. They offer greater availability to a broad range of users. The adaptability of multimodal interfaces is apparent in their capability to switch input modes when situations and environment warrant. The simultaneous input possibilities they provide allow for more efficient input, and the ability of multimodal systems to use mutual disambiguation is an advantage that facilitates error avoidance and recovery.

5. REFERENCES

[1] Christopher A. Robbins. ,2004 Speech and Gesture Based Multimodal Interface Design

[2] Joyce Y. Chai, MIND: A Context-Based Multimodal Interpretation Framework In Conversational Systems

[3] Marcelo Worsley And Michael Johnston " Multimodal Interactive Spaces: Magictv And Magicmap "IEEE Vol 978-1-4244-7903-2010.

[4] Meng and Yuyang Zhang and Yaochu Jin "Autonomous Self –Reconfiguration of Modular Robots by evolving a Heirarchical Model" 2011.

[5] Boucher, R. Canal, T.-Q. Chu, A. Drogoul, B. Gaudou, V.T. Le, V.Moraru, N. Van Nguyen, Q.A.N. Vu, P. Taillandier, F. Sempe, and S. Stinckwich. :"A Real- Time Hand Gesture System based on Evolutionary Search".In Safety, Security Rescue Robotics (SSRR), 2011 IEEE International Workshop on, pages 16, 2011.

[6] Rami Abielmona ,Emilm.Petriu, Moufid Harb and Slawo Ye solkowki,"Mission Diven Robotics for Territorial Security" Model"IEEE transaction on Computational Intelligence Magzine ,pp 55-67 Feb 2011.

[7] Surakka, Martti Juhola, and Jukka Lekkal "A Wearable, Wireless Gaze Tracker with Integrated Selection Command Source for Human–Computer Interaction ", IEEE transaction on Computational Intelligence Magzine , 2011.

[8] Sharon Oviatt and Phil Cohen ,Lizhong Wu ,John Vergo Lisbeth Duncan ,Bernhard Suhm and Josh Bers ,Thomas Holzman ,Terry Winograd ,James Landay ,Jim Larson David Ferro , Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions, HUMAN-COMPUTER INTERACTION, 2000, Volume 15, pp. 263–322

[9] Joyce Chai, Shimei Pan, Michelle X. Zhou and Keith Houck," Context-based Multimodal Input Understanding in Conversational Systems", Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02) ,0-7695-1834-6/02 $17.00 © 2002 IEEE