Intelligent Multimodal Systems

Sharda.A.Chhabria
Sr. Lecturer, IT
Sharda_chhabria@yahoo.co.in

G.H.Raisoni College of Engg.

Dr.R.V.Daraskar
Professor, CSE
rvdharaskar@yahoo.com
rvdharaskar@rediffmail.com
G.H.Raisoni College Of Engg.

Abstract.

The field of human-computer interaction has seen innumerous advances and innovations in the last few years. Developments in both hardware and software enable the use of speech, gestures, body posture, different tracking technology, tactile/force feedback devices, eye-gaze and even bio-sensors to develop a new generation of applications. In a multimodal conversation, the way users communicate with a system depends on the available interaction channels and the situated context (e.g., conversation focus, visual feedback). A correct interpretation can only be attained by simultaneously considering these constraints. Here, we present an overview of the technologies integrated to build multimodal systems, review the various scenarios in which multimodal systems have been incorporated by researchers. This is literature survey research paper which focuses on developing intelligent multimodal user interface using more efficient & optimized algorithms through eye, speech, and gestures as inputs.

Keywords*: Multimodal interface. input modes -- e.g., voice, text, mouse clicks, and gestures -- and output modes -- voice, text and graphics,*

Introduction .

Intelligent multi-modal systems use a number of input or output modalities to communicate with the user, exhibiting some form of intelligent behaviour in a particular domain. [1] The functional requirements of such systems include the ability to receive and process user input in various forms such as:

- typed text from keyboard,
- mouse movement or clicking,
- speech from a microphone,
- focus of attention of human eye captured by a camera,

The system must be also able to generate output for the user using speech, graphics, and text.A system, which exhibits the above features, is called a multi-modal system. For a multi-modal system to be also called intelligent, it should be capable of reasoning in a particular domain automating human tasks, facilitating humans to perform tasks more complex than before or exhibiting a behaviour which can be characterized as intelligent by the users of the system.

Given these requirements for intelligent multi-modal systems, it becomes obvious that such systems, in general, are difficult to develop. A modular approach is therefore necessary for breaking down the required functionality into a number of sub-systems which are easier to develop or for which software solutions already exist. Other requirements for such systems are concurrency, a communication mechanism and distribution of processes across a network.

Today's Technology.

In our increasingly technological society, we start already finding multi-modality supporting devices in daily life; from voice-controlled cell phones to personal digital assistance (PDA) supporting some form of handwriting recognition. While such already habitual human-machine interfaces improve effective interaction in and within the real world, they still suffer from the same restrictions as their counterpart desktop applications: the interfaces' lack of awareness with respect to the current user and application environment. Such awareness is even more crucial since devices as the ones described above, are typically used in a multitude of environments by a wide range of people.

Video tracking together with image processing presents an approach to detect user, presence, capture gestures and postures. It also resolves some of the restrictions imposed by the current clumsy and *tethered* tracking systems and seems a promising solution for mobile systems. The computer gathers knowledge about the user's attention through eye gaze tracking devices; wearable physiological monitoring systems have been built to acquire users' emotional expressions.

These are just some examples of technologies that address the user and application environment awareness. Fortunately, the exponential growth of computer performance starts to produce devices to drive some of these systems that literally fit into our hands.

An important factor for the relatively wide spread of technology is posed by the economical factor. The constant drop of prices increases accessibility of technology to new markets. As an example, it is becoming more common for elementary school students to carry a cell phone. Cell phones have a multitude of functions not necessarily useful for children, so cell phone companies created fun looking and simpler versions for this specific market.

The bottom line here is: different users with different needs are presented with different devices. Such approach may work with some markets but it is not feasible or economically viable for every market, particularly if we start to address small groups of people with specific impedances.

Let's think what happens nowadays with books for vision impaired people, how long it takes for a book to be edited with larger fonts or in braille? A book though an excellent media for reading cannot change its content, but that's precisely what digital devices is good at.

The current state-of-the-art research of human-computer interaction gives a first approach to the solutions for the questions above. Video tracking together with image processing presents an approach to detect user presence, capture gestures and postures. It also resolves some of the restrictions imposed by the current clumsy and *tethered* tracking systems and seems a promising solution for mobile systems. The computer gathers knowledge about the user's attention through eye gaze tracking devices; wearable physiological monitoring systems have been built to acquire users' emotional expressions. These are just some examples of technologies that address the user and application environment awareness. Fortunately, the exponential growth of computer performance starts to produce devices to drive some of these systems that literally fit.

Research Challenges.

Without doubt the obtrusiveness presented by these technologies imposes a main hindrance to universal accessibility. Though looking at the evolution in hardware over the last years it is predictable that technological advances will make the use of these devices less cumbersome and robust and it will not be too long before they can be employed within our daily life. On the other hand, it is certainly also true that the obstacles are not only at the technological level. Increasing the amount of the information about the
user or his environment is of little or no use if there is no knowledge available on how to process it. The knowledge on how the human-computer communication modalities can be used in conjunction to improve the effectiveness of human-computer interfaces is far from being complete.

Multimodal applications Challenges

Combining text and speech in the same application is an example of a "multimodal" interface. Multimodal interfaces combine different input modes -- e.g., voice, text, mouse clicks, and gestures -- and output modes -- voice, text and graphics, and perhaps others. The user can switch back and forth between modes during the application, at one point speaking a choice, at another point gesturing with a stylus.

Multimodal applications present fascinating challenges to user interface designers:

- With various modes all active simultaneously, the user interface (UI) must present a consistent conceptual model to the user and provide consistent information.
- If speech recognition is one of the input modes, the other modes can provide enormous help by prompting the user to speak an easily-recognized utterance.

The new mobile displays create a demand for an efficient technique to interact with the information displayed in the HMD. When the hands are needed for other tasks, hand-controlled devices such as keyboard or mouse become awkward. Gaze interaction with the HMD can potentially provide a hands-free pointing technique.
People using augmented and alternative communication tools may benefit from an HMD with gaze control. Daily activities, like driving a wheelchair, would not be interrupted when communicating. People without control of their hands could communicate on-the-move and in bed without requiring external assistance to reposition the equipment.

Related Work.

Research in human-computer interactions [1] has mainly focused on natural language, text, speech and vision primarily in isolation. Recently there have been a number of research projects that have concentrated on the integration of such modalities using intelligent reasoners .The rationale is that many inherent ambiguities in single modes of communication can be resolved if extra information is available .Among the projects reviewed in the references are CUBRICON from Calspan-UB Research Centre, XTRA from German Research Centre for AI and the SRI system from SRI International.

Previously, [12] work in the field of gesture recognition usually first segmented parts of the input images -- for example the hand -- and then used features calculated from this segmented input. Results in the field of object recognition in images suggest that this intermediate segmentation step is not necessary and we can instead employ features directly obtained from the input images, so-called appearance-based features. In this work, we show that using these features and appropriate models of image variability, we can obtain excellent results for gesture recognition tasks. Very good results can be obtained using a downscaled image of each video frame and tangent distance as a model of image variability. Also a new dynamic tracking algorithm is introduced which makes its tracking decisions at the end of a video sequence using the information of all frames. This tracking method allows for tracking under very noisy circumstances.[12]

In the last decade, [13] the growth and the popularity of the World Wide Web (Web) have been phenomenal. Originally, it was a purely text-based system that allowed assistive technologies to be designed to transform pages into alternative forms (e.g., audio) for disabled people. This meant that for the first time, a vast amount of information was available and easily accessible to disabled people. However, advances in technologies and changes in the main authoring language, transformed the Web into a true visual communication medium. These changes eventually made the Web inaccessible to visually impaired users. In particular, travelling around the Web became a complicated task, since the richness of visual navigational objects presented to their sighted counterparts are neither appropriate nor accessible to visually impaired users.

This thesis investigates principles and derived techniques to enhance the travel experience for visually impaired Web users. The hypothesis is that travel support for visually impaired users can be improved if Web pages are analysed to identify the objects that support travel and are then transformed in such a way that they can then fulfill their intended or implicit roles. This hypothesis is supported by the identification of structural and navigational properties of Web pages which have been encapsulated into ontology (WAfA) to support machine processing; and the design of a flexible pipeline approach to annotate and transform Web pages by using this ontology. An experimental test-bed, Dante, has also been created based on this pipeline approach that encodes these principles and techniques to transform Web pages. Finally, a user evaluation method is devised and applied to demonstrate that the travel experience of visually disabled users can be improved through the application of these techniques.

This research demonstrates that by providing machine processable data, regarding the structural and navigational properties of Web pages, applications can be created to present Web pages in alternative forms and so enhance the travel experience of visually impaired users. The work presented in this thesis is of practical value to the Web accessibility community and is an important case study for demonstrating Semantic Web technologies. By moving away from thinking that simple translation of text to audio is enough to provide access to

Web pages, this thesis is part of the coming of age of assistive technology and is a significant contribution to redressing the inequality arising from visual dominance in the Web.[13].

Proposed Work.

This research plans on developing multimodal user interface using more efficient & optimized algorithms through eye, speech, and gestures as inputs.

- This modal will be helpful for disabled persons.
- It will perform various operations like typing through eye or speech, mouse operations etc.
- It will try to cover different disabilities for different type of persons.
- Also a unique point of this research may be that interface will work for different languages spoken by different people.
- With various modes all active simultaneously, the user interface (UI) will present a consistent conceptual model to the user and provide consistent information.
- If speech recognition is one of the input modes, the other modes will provide enormous help by prompting the user to speak an easily-recognized utterance.
- Thus the overall system will also be helpful to normal people as it will be fast & user friendly.

Brief References.

1] Azvine, B., Azarmi, N. and Tsui, K.C. 1996. Soft computing - a tools for building intelligent Systems, BT Technology Journal, vol. 14, no. 4, pp. 37 45, October.

2] Avatar Mediated Conversational Interfaces R. Alex Colburn, Michael F. Cohen, Steven M. Drucker,7/31/2000,Technical Report MSR-TR-2000-81, Microsoft Research, Microsoft Corporation, One Microsoft Way

3] A Salience Driven Approach to Robust Input Interpretation in Multimodal Conversational Systems Joyce Y. Chai Shaolin Qu ,Computer Science and Engineering, Michigan State University, East Lansing, MI 48824.

4] Chai, J., Pan, S., Zhou, M., and Houck, K. Context-based Multimodal Interpretation in Conversational Systems. *Fourth International Conference on Multimodal Interfaces*, 2002.

5] An Exploration of Eye Gaze in Spoken Language Processing for Multimodal Conversational Interfaces ,Shaolin Qu Joyce Y. Chai ,Department of Computer Science and Engineering,Michigan State University, East Lansing, MI 48824 Proceedings of NAACL HLT 2007, pages 284–291,Rochester, NY, April 2007. c2007 Association for Computational Linguistics.

6] Eye-movements and Voice as Interface Modalities to Computer Systems Mohsen M. Faridand Fionn Murtagh School of Computer Science, Queen's University Belfast. Belfast, Northern Ireland, BT7 1NN, United Kingdom.

7] Predicting User Attention using Eye Gaze in Conversational Interfaces Zahar Prasov ,Joyce Chai, Department of Computer Science Department of Computer Science Michigan State University Michigan State University, East Lansing, MI 48823 East Lansing,

8] Cognitive Principles in Robust Multimodal Interpretation Joyce Y. Chai, Zahar Prasov Shaolin Qu**,** Department *of Computer Science and Engineering Michigan State University, East Lansing, MI 48824 USA.,*c 2006 AI Access Foundation.

9] Proceedings of COGAIN 2008, 'Communication, Environment and Mobility Control by Gaze' The 4th Conference on Communication by Gaze Interaction – COGAIN 2008: Communication, Environment and Mobility Control by Gaze, September 2-3, 2008 1, Prague, Czech Republic

10] Using Eye Gaze Patterns to Identify User Tasks, Shamsi T. Iqbal and Brian P. Bailey Department of Computer Science, University of Illinois, Urbana, IL 61801, USA.

11] Multimodal Interaction:an integrated speech and gaze approach Relatore: Candidata:,Prof. Fulvio Corno Alessandra Pireddu Corelatore:Ing. Laura arinetti,Aprile 2007

12] Appearance-Based Gesture Recognitions, Philippe Drew, Advisor: Dr. Daniel Keysers , University of Colorado, RWTH Aachen University Germany, January 2005

13] PhD Thesis: Annotation and Transformation of Web Pages to Improve Mobility for Visually Impaired Users Yeliz Yesilada, Advisors: Carole Goble and Robert Stevens, The University of Manchester, United Kingdom ,2005.