Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speech-Gesture Interpretation System using Human Computer Interface

M.V. Lande
M.Tech. CSE
PCST ,Indore
E-mail:milind.jdiet@gmail.com

Prof.Makarand Samvatsar
Computer Sci &Engg, Dept.
PCST,Indore

Mr. A.G.Barsagade
I.T.Department
RCERT Chandrapur

ABSTRACT: **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. The different computational approaches that may be applied at the different levels of modality integration. Thus this system is needed for interpreting and fusing multiple sensing modalities in the context of 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 interface, Speech, multimodal interface.

1. INTRODUCTION

Gesture Interpretation is a topic in science and technology with the goal of interpreting human gestures via mathematical algorithms. Gestures can originate from any bodily motion or state but commonly originate from the face or hand. Current focuses in the field include emotion recognition from the face and hand gesture recognition. Many approaches have been made using cameras and computer vision algorithms to interpret sign language. However, the identification and recognition of posture, gait, proxemics, and human behaviors is also the subject of gesture recognition techniques Gesture Interpretation enables humans to interface with the machine (HMI) and interact naturally without any mechanical devices. Using the concept of gesture recognition, it is possible to point a finger at the computer screen so that the cursor will move accordingly. This could potentially make conventional input devices such as mouse, keyboards and even touch-screens redundant. Recent advances in various technologies, coupled with an explosion in the available computing power, have given rise to a number of novel human–computer interaction modalities-speech, vision-based gesture recognition, eye tracking, electroencephalograph, etc. Successful embodiment of these modalities into an interface has the potential of easing the human computer interface bottleneck that has become noticeable with the advances in computing and communication.

Now the hands-free phone in a car relies on computing devices that react to our spoken words. New wireless mobile devices are used to transfer money from our bank accounts through the touch of fingertips. Just as these ideas were once considered fanciful, technologies expected to be commercialized over the next several years have possibilities, if we seize them, which could serve the engineering profession, society, and the economy in ways that are impossible now. It is the purpose of all technologies to improve the quality of life, and of work, as well. Technology is intended to make things better-that is, safer, easier, more satisfying, and therefore more enjoyable. It extends what people are capable of achieving.

Significant effort is currently being devoted to making human interactions with computers, physical systems, and with information in general, simple, natural, and seamless. The objectives of many of the recent developments are to enhance productivity and accelerate innovation. The pace of advances in computing, communication, mobile, robotic, and interactive technologies is accelerating.

2. LITERATURE REVIEW

Face-to-face communication is highly interactive. Even when only one person speaks at the time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. Such feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participants ability to communicate [3, 13]. It argues that it is possible to significantly improve state-of-the art recognition techniques by exploiting regularities in how people communicate. People do not provide feedback at random. Rather they react to the current topic, previous utterances and the speaker's current verbal and nonverbal behavior [1]. For example, listeners are far more likely to nod or shake if the speaker has just asked them a question, and incorporating such dialogue context can improve recognition performance during human-robot interaction [4].

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. [OVIATT02]. Examples of such types of interfaces include CUBRICON [NEAL90], XTRA [WAHLSTER91], and Shoptalk [COHEN92].

2.1. Put-That-There

In Bolt ‘s —Put-That-There” system, speech recognition is used in parallel with gesture recognition. 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 [BOLT8]. Speech recognition in the —Put That There” system allows for simple English sentence structures using a limited vocabulary.

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 [NEAL9], 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.

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. 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.

3. MOTIVATION

Gesture recognition systems identify human gestures and the information they convey. Although relying on gesture as the primary source of command input to computers may sound like science fiction, the technology has rapidly progressed in some areas such as virtual reality, by relying on special hardware and wearable devices . This hardware is often not cost-effective and is infeasible for some applications; consequently, gesture recognition based on alternative methods of Data Acquisition is being considered. In this article we introduce A novel Method for gesture Recognition.

4. SYSTEM ARCHITECTURE

Humans perceive the environment in which they live through their senses—vision, hearing, touch, smell, and taste. They act on and in it using their actuators such as body, hands, face, and voice. Human-to-human interaction is based on sensory perception of actuator actions of one human by another, often in the context of an environment. In the case of human computer interface, computers perceive actions of humans. To have the human– computer interaction be as natural as possible, it is desirable that computers be able to interpret all natural human actions. Hence, computers should interpret human hand, body, and facial gestures, human speech, eye gaze, etc. Some computer-sensory modalities are analogous to human ones as shown in following fig 4.1(a),4.1(b)

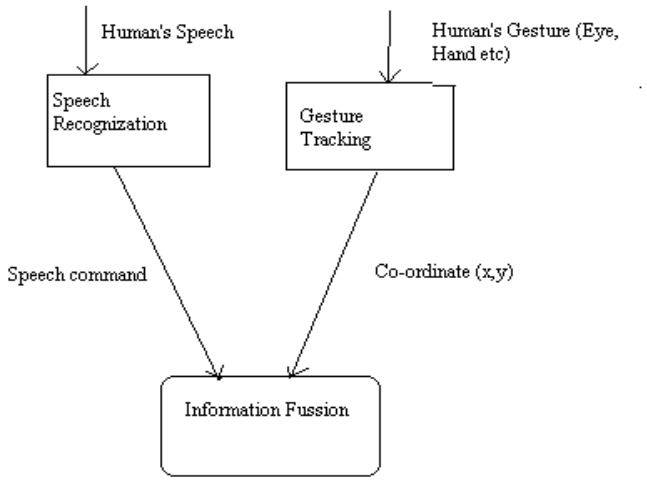

Fig. 4.1(a)

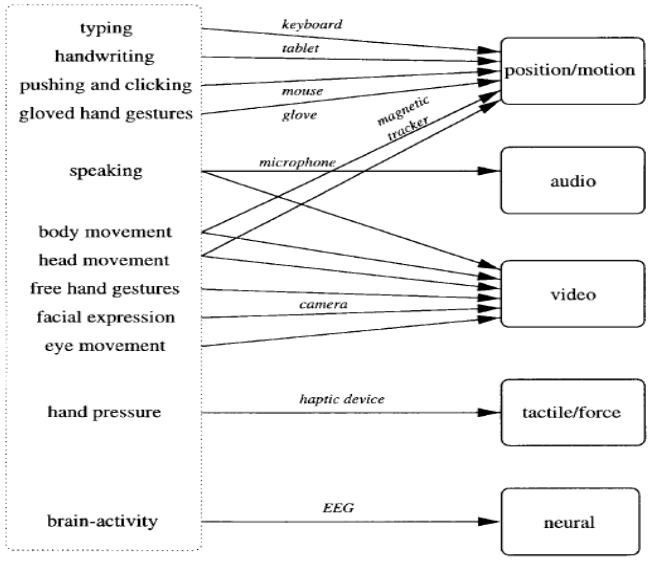

Fig. 4.1(b): Mapping of different human-action modalities to computer-sensing modalities for Human Computer Interface.

Computer vision and Automatic Speech Recognization mimic the equivalent human sensing modalities. However, computers also possess sensory modalities that humans lack. They can accurately estimate the position of the human hand through magnetic sensors and measure subtle changes of the electric activity in the human brain, for instance. Thus, there is a vast repertoire of human-action modalities that can potentially be perceived by a computer. Multiple human actions, such as facial expressions and hand or eye movement, can be sensed through the same ―devices ‖ and used to infer different information. The modalities are discussed under the two categories of human-action modalities and compute sensing modalities. A particular human-action modality (e.g., speaking) may be interpreted using more than one computer-sensing modality (e.g., audio and video).

The action modalities most exploited for gesture interpretation system are based on hand movements. This is largely due to the dexterity of the human hand which allows accurate selection and positioning of mechanical devices with the help of visual feedback. Appropriate force and acceleration can also be applied easily using the human hand. Thus, the hand movement is exploited in the design of numerous interface devices—keyboard, mouse, stylus, pen, wand, joystick, trackball, etc. The keyboard provides a direct way of providing text input to the computer, but the speed is obviously limited and can only be improved to a certain rate. Similarly, hand movements cause a cursor to move on the computer screen (or a 3-D display). The next level of action modalities involves the use of hand gestures, ranging from simple pointing through manipulative gestures to more complex symbolic gestures such as those based on American Sign Language. With a glove-based device, the ease of hand gestures may be limited, but with non-contact video cameras, free-hand gestures would be easier to use for Gesture Interpretation System. The role of free-hand gestures in Gesture Interpretation System could be further improved (requiring lesser training, etc.) by studying the role of gestures in human communication. A multimodal framework is particularly well suited for embodiment of hand gestures into human computer interface.

In addition to hand movements, a dominant action modality in human communication is the production of sound, particularly spoken words. The production of speech is usually accompanied by other visible actions, such as lip movement, which can be exploited in Gesture Interpretation System as well. Where the human is looking can provide a clue to the intended meaning of a particular action or even serve as a way of controlling a display.

Thus, eye movements can be considered a potential action modality for Gesture Interpretation System. The facial expression and body motion, if interpreted appropriately, can help in human computer interface. Even a subtle- action" like a controlled thought has been investigated as a potential candidate for human computer interface.

5. ISSUES IN DESIGNING GESTURE INTERPRETATION SYSTEM

In this section, we outline both the scientific and engineering challenges in designing speech–gesture driven multimodal interfaces in the context based gesture interpretation system. Our main goal is to design a dialogue-enabled HCI system for collaborative decision making, command, and control. While traditional interfaces support sequential and unambiguous input from devices such as keyboard and conventional pointing devices (e.g., mouse, trackpad), speech–gesture driven dialogue-based multimodal interfaces relax these constraints and typically incorporate a broader range of input devices (e.g., spoken language, eye and head tracking, speech, gesture, pen, touch screen, displays, keypads, pointing devices, and tactile sensors). The ability to develop a dialogue-based speech–gesture driven interface is motivated by the knowledge of the natural integration patterns that typify people‘s combined use of different modalities for natural communications. Recent trends in multimodal interfaces are inspired by goals to support more transparent, flexible, efficient, and powerfully expressive means of HCI than ever before. Multimodal interfaces are expected to support a wider range of diverse applications, to be usable by a broader spectrum of the average population, and to function more reliably under realistic and challenging usage conditions. The main challenges related to the design of a speech–gesture driven interface for gesture interpretation system are:

1. domain and task analysis;
2. acquisition of valid multimodal data;
3. speech recognization;
4. recognizing users gesture;
5. a framework to fuse gestures and spoken words;
6. interoperability of devices.

We next discuss each of these challenges in some detail.

5.1. Domain and Task Analysis

Understanding the task domain is essential to make the challenge of building a natural interface for gesture interpretation system (or other application domains) a tractable problem. This is because multimodal signification (through speech, gesture, and other modalities) is context dependent. Within this context, cognitive systems engineering (CSE) has proven to be an effective methodology for understanding the task domain and developing interface technologies to support performance of tasks [14]–[15]. The theoretical frameworks of distributed cognition [7], activity theory [8], and cognitive ergonomics [11] also have the potential to help isolate and augment specific elements of the crisis management domain for multimodal system design. one should consider scale and needs before settling on a single framework, making it important to consider a variety of approaches in designing a collaborative multimodal gesture interpretation System.

5.2. Acquisition of valid multimodal data

An important feature of a natural interface would be the absence of predefined speech and gesture commands. The resulting multimodal ―language‖ thus would have to be interpreted by a computer. While some progress has been made in the natural language processing of speech, there has been very little progress in the understanding of multimodal HCI [14]. Although, most gestures are closely linked to speech, they still present meaning in a fundamentally different form from speech. Studies in human-to-human communication, psycholinguistics, and others have already generated a significant body of research on multimodal communication. However, they usually consider a different granularity of the problem. The patterns from face-to-face communication do not automatically transfer over to HCI due to the ―artificial‖ paradigms of information displays. Hence, the lack of multimodal data, which is required to learn the multimodal pattern, prior the system building creates so-called chicken-and-egg problem.

4.3. Speech Recognization

Speech recognition requires the computer to accept spoken words as input and interpret what has been spoken. To make the job of understanding speech easier for the computer, a method of speech input called command and control is used. Speech Recognition is technology that allows a computer to identify the words that a person speaks into a microphone. We used Microsoft Agent version 2.0 that provides a library for more natural ways for people to communicate with their computers. And also we used The Lernout & Hauspie TruVoice Text-to-Speech (TTS) [18] Engine that provides speech output capabilities for Microsoft Agent so we can hear what the characters are saying through your sound speakers.

The commands available to the user are the following: “left click” (or just “click”), “double left click” (or just “double click”), “right click”, “double right click”. The user can also keep a button pressed so as to highlight a group of objects. The command “down”, “up” change the selection area of the mouse per example in the menu (File, Edit, Insert…) and we can select also the menu with the voice command. This set of available commands allows executing meaningful tasks on the computer since all the main mouse click operations are available.

Improving performances in voice recognition can be done taking into account the following criteria:
- dimension of recognizable vocabulary;
- spontaneous ness degree of speaking to be recognized;
- dependence / independence on the speaker;
- time to put in motion the system
- system accommodating time at new speakers;
- decision and recognition time;
- recognition rate (expressed by word or by sentence).

Today's vocal recognition systems are based on the general principles of forms' recognition[3][7]. The methods and algorithms that have been used so far can be divided into four large classes:

- Discriminate Analysis Methods based on Bayesian discrimination;
- Hidden Markov Models;
- Dynamic Programming-Dynamic Time Warping algorithm [8];
- Neuronal Networks.

5.4. Recognizing User's Gesture

Gesture acquisition is concerned with the capture of the hand/body motion information in order to perform subsequent gesture recognition. Gestures are in general defined as movement of the body or limbs that expresses or emphasizes ideas and concept. In the context of multimodal systems, penand touch-based interfaces are also commonly viewed to fall under the gesture recognition domain. However, while for pen- and touch-based systems, gesture acquisition is merely a marginal problem, it requires considerable effort for most other approaches. Aside from pen- and touch-based systems [9], the most common gesture acquisition methods are based on magnetic trackers, cyber-gloves and vision-based approaches. The suitability of the different approaches depends on the application domain and the platform. Pen based approaches are the method of choice for small mobile devices and are cost effective and reliable. Acquisition using magnetic trackers [12] and/or cyber gloves is efficient and accurate but suffers from the constraint of having to wear restrictive devices. In contrast, vision-based approaches offer entirely contact-free interaction and are flexible enough to operate on all platforms except the smallest mobile devices.

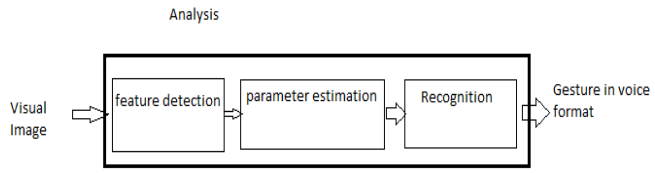

Fig.5.4.1 Analysis and Recognition of Gestures

5.5. A Framework to Fuse Gestures and Spoken Words:

We illustrate the architecture of a possible fusion strategy. We believe that a probabilistic evaluation of all possible speech– gesture combinations promises a better estimation of users intent than either modality alone. The conditional probabilities of observing certain gestures given a speech utterance will be based on several factors. Speech utterances will first have to be analyzed for keyword classes such as typical deictic keywords (e.g., this,"that"). These keywords can then be associated with corresponding deictic gestures. The association needs to take gesture and utterance component classes into consideration and maintain the appropriate mapping between speech and gesture components.

Once data associations (or set of associations if several are possible) have been determined, the co-occurrence module can determine a final match value between the utterance and the gesture based on temporal co-occurrence statistics.

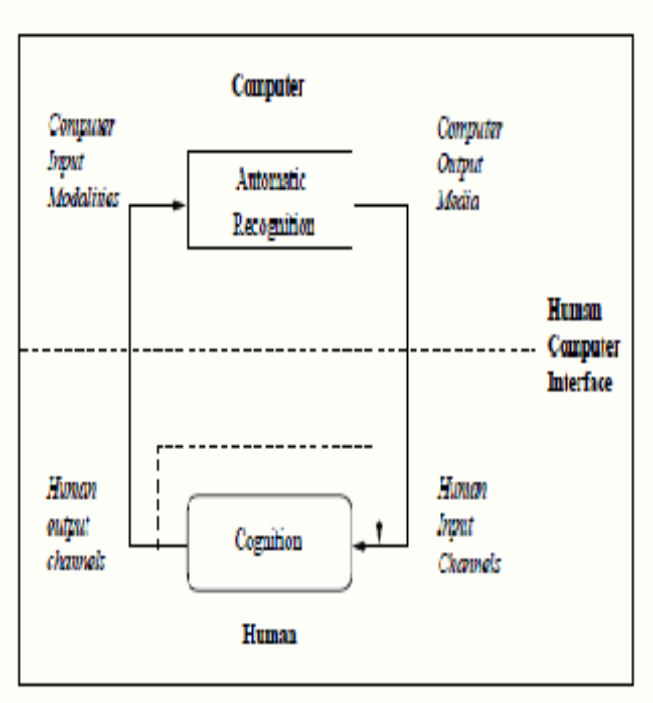

Fig 5.5.1: A Framework Architecture

5.6. Interoperability of devices

Both the interpretation of multimodal input and the generation natural and consistent responses require access to higher level knowledge. In general, semantics required by multimodal systems can be categorized along two dimensions: general versus task/domain specific, and dynamic versus staticoccurrence relations between speech and gesture. The issue of interoperability across the wide range of devices is very critical for a seamless flow of information and communication. Hence, it is important to design unified multimedia applications

6. CONCLUSION

In gesture interpretation system Human-Computer Interaction is an important part of systems design. Quality of system depends on how it is represented and used by users. Therefore, enormous amount of attention has been paid to better designs of HCI. The new direction of research is to replace common regular methods of interaction with Intelligent ,adaptive, multimodal, natural methods. Motivated by the tremendous need to explore better HCI paradigms,

there has been a growing interest in developing novel sensing modalities for HCI. To achieve the desired robustness of the HCI, multimodality would perhaps be an essential element of such interaction. Clearly, human studies in the context of HCI should play a larger role in addressing issues of multimodal integration. Even though a number of developed multimodal interfaces seem to be domain specific, there should be more systematic means of evaluating them. Modeling and computational techniques from more established areas such as sensor fusion may shed some light on how systematically to integrate the multiple modalities. However, the integration of modalities in the context of HCI is quite specific and needs to be more closely tied with subjective elements of -context. There have been many successful demonstrations of HCI systems exhibiting multimodality. Despite the current progress, with many problems still open, multimodal HCI remains in its infancy. A massive effort is perhaps needed before one can build practical multimodal HCI systems approaching the naturalness of human–human communication.

FUTURE IMPROVEMENTS

1. After improving Scaling, PCA with different distance from camera can be analyzed and if accuracy is high, then can be implemented in real time with video implementation.
2. Training of ANN: Analyzing with different number of layers and different number of neurons per layer, a good trained model can be obtained. This is expected to be more Robust than current hierarchical classification.
3. Extracting more features with Multiple View Point concept, as there are more data points to analyze per histogram or, we can say we have multiple signatures for each gesture and extracting different features from each signature.
4. Hierarchical approach can be implemented together with PCA or, with ANN for better results. This can be in a way that first we find number of fingers in the gesture from which we can have 6 classes and with 6 (1 for each class) different Eigen Spaces for PCA we can expect better results.

7. REFERENCES

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

[2] Rajeev Sharma, Mohammed Yeasin, Member, Ieee, Nils Rahnstoever, Ingmar Rauschert, Guoray Cai, Member, Ieee, Isaac Brewer, Alan M. Maceachren, And Kuntal Sengupta, -Speech–Gesture Driven Multimodal Interfaces for Crisis Management"Proc Of The Ieee, Vol. 91, No. 9, September 2003.

[3] S.A. Chhabria and R.V. Dharaskar,-Multimodal interface for disabled persons"in International Journal of Computer Science and Communication ,2011.

[4] 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.

[5] M. Segers, James Connan, -Real-Time Gesture Recognition using Eigenvectors" Vaughn Private Bag X17 Bellville, 7535, volume III,2009.

[6] Rami Abielmona ,Emilm.Petriu, Moufid Harb and Slawo Yesolkowki,"Mission Diven Robotics for Territorial Security"Model"IEEE transaction on Computational Intelligence Magzine ,pp 55-67 Feb 2011. 36

[7] Melody Moh, Benjamin Culpepper,Lang Daga, -Computer Vision and Pattern Recognition -IEEE , 2005. CVPRW -05.Conference on, page 158, June 2005.

[8] Boukje Habets, Sotaro Kita, Zeshu Shao, Asli Özyurek, and Peter Hagoort -The Role of Synchrony and Ambiguity in Speech-Gesture Integration during Comprehension" 2011.

[9] R. Sharma, V. I. Pavlovic, and T. S. Huang,-Toward multimodal human-computer interface," Proc. IEEE, vol. 86, pp. 853 – 869, May 1998.

[10] R. Stiefelhagen, C. F ¨ ugen, P. Gieselmann, H. Holzapfel, K. Nickel and A. Waibel -Natural Human-Robot Interaction using Speech, Head Pose and Gestures"Proceedings of the Third IEEE International Conference on Humanoid Robots - Humanoids 2003.

[11] Benoit Legrand, C.S. Chang, S.H. Ong, Soek-Ying Neo, Nallasivam Palanisamy, -Chromosome classification using dynamic time warping", ScienceDirect Pattern Recognition Letters 29Dec 2008.

[12] Mohammad Hasanuzzaman, Saifuddin Mohammad Tareeq, Vuthichai Ampornaramveth, Hironobu Gotoda-Adaptive Visual Gesture Recognition For Human-Robot Interaction" Malaysian Journal Of Computer Science, Vol. 20(1), 2007

[13] Ville Rantanen, Toni Vanhala, Outi Tuisku, Pekka-Henrik Niemenlehto, Jarmo Verho, Veikko 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 .

[14] Marcelo Worsley And Michael Johnston " Multimodal Interactive Spaces: Magictv And Magicmap " IEEEVol 978-1-4244-7903-2010

[15] Y. Tamura, M. Sugi, J. Ota, and T. Arai, -Estimation of user ' s intention inherent in the movements of hand and eyes for the deskwork support system," in IEEE/RSJ IROS, (USA), pp. 3709-3714, Nov. 2007.

