A REVIEW PAPER ON MULTIMODAL HUMAN-COMPUTER INTERACTION

[1]Fatema vhora, [2]Amita Tailor, [3]Jay Gandhi , [4]Urvashi Parmar [5]Nirali Bhaliya [1]Lecturer, [2]Lecturer , [3]Asst. Professor , [4]Lecturer, [5]Student

[1]CSE department,

[1]Parul University, Vadodara, India

Abstract : People naturally interact with the world multimodally, through both parallel and sequential use of multiple perceptual modalities. Multimodal human–computer interaction has sought for decades to endow computers with similar capabilities, in order to provide more natural, powerful, and compelling interactive experiences. With the rapid advance in non-desktop computing generated by powerful mobile devices and affordable sensors in recent years, multimodal research that leverages speech, touch, vision, and gesture is on the rise. This paper provides a brief and personal review of different types of multimodal like gaze tracking and keystroke, how it is useful in augmented reality, games, 3d images, child-robot interaction etc. Finally, we list challenges that lie ahead for research in multimodal human–computer interaction

Keywords **- Multimodal human interaction, gaze tracking, augmented reality, 3d images, games, child-robot interaction**

I. INTRODUCTION

Recent years many different areas including computer vision, Artificial intelligence, psychology, and many others are using Multimodal Human computer interaction (MMHCI). Because computer is interacted as an object with human in many different ways using different applications so, user needs to be able to interact with computer naturally like face-to-face human-human interaction. Generally we communicate through speech and body language (vision, posture, gaze, hand motion) to express emotion, mood, attitude, and attention[1]. Multimodal human-computer interaction refers to the "interaction with the virtual and physical environment through natural modes of communication". Multimodal systems can offer a flexible, efficient and usable environment allowing users to interact through input modalities, such as speech, handwriting, hand gesture and gaze, and to receive information by the system through output modalities, such as speech synthesis, smart graphics and other modalities, opportunely combined. Then a multimodal system has to recognize the inputs from the different modalities combining them according to temporal and contextual constraints in order to allow their interpretation. This process is known as multimodal fusion, and it is the object of several research works. In this paper we focus on gaze tracking, immersive system using multimodal, and multimodal in augmented reality, multimodal child interaction, natural interface for human drone, Multimodal Mixed Reality Real-Time Strategy, Game, emotion recognition, natural interaction with 3D image, impact of keyboard design. All these factors affect in multimodal implementation.

II MOTIVATION

Multimodal interaction provides the user with multiple modes of interacting with a system. A multimodal interface provides several distinct tools for input and output of data. For example, a multimodal question answering system employs multiple modalities (such as text and photo) at both question (input) and answer (output) level. So, after reading this paper one can able to understand how interaction takes place between human-computer and how it is useful for our day to day life.

III LITERATURE SURVEY

Multimodal human computer interaction including much research area such as computer vision, psychology, artificial intelligence etc. The main focus on problem of vision and this problem can be solved using multimodal human computer interaction. Group vision techniques according to the human body movement, gesture like hand and gaze analysis are used for tasks as emotion recognition in affective interaction and for a variety of applications. Here discuss the affective computer interaction issues in multi-modal fusion, modeling, and data collection and a variety of emerging multimodal human computer interaction application. Multimodal human computer interaction is a very dynamic and broad research area we do not intend to present a complete survey.[1]

The Digital interactive surfaces augment traditional physical setups with virtual elements and touch-based interactions to create a Mixed Reality. Such environments overcome the static nature of traditional multi-user tabletop games. In this paper, introduce a new combination of interaction styles and game play mechanics for mixed reality-based tabletop role-playing games. Here 3 approaches is very important 1) Leave the traditional turn-based style in favor of an increased simultaneous real-time interaction 2) based on intelligent virtual agents 3) controlled via a multimodal speech and gesture interface. Such interfaces provide interaction alternatives given the dynamic nature of many applications in MR as well as in Virtual and Augmented Reality [2]

A survey on theoretical and practical work of the latest research in emotion recognition from multi-modal information including facial and vocal expression. In this paper, Two main issues in the current research on automatic analysis of facial expressions consider facial affect (e.g., emotion) recognition and facial muscle action recognition .Most facial affect recognition issue attempts to recognize a small set of prototypical emotional facial expressions such as the six prototypical emotions: anger, disgust, fear, happiness, sadness and surprise. Facial Action Coding System (FACS) was proposed to classify the atomic facial signals into Action Units (AUs) through analysis of facial muscle contractions. Using FACS, human coders can manually code nearly any anatomically possible facial expression, and the us can be used as the basis for any higher order decision making process including the recognition of prototypical emotions, cognitive states, social signals, and more. FACS motivated the studies

on automatic human spontaneous facial behavior recognition. Developing a better data fusion approach is desirable obtain the advantages of various fusion strategies, such as combining the model-level and decision-level fusion properties [3]

In this Paper, Controlling three dimensional images with gestures and speech using a three dimensional depth camera is realized in order to ensure multi-modal natural interaction with computers. Realized system allows touch less interaction by using speech and gesture recognition rather than the use of keyboard or mouse. In the system starting and closing operations are conducted by speech commands while interaction with 3-D images is done by gestures. Microsoft Kinect device was chosen for the implementation of the system since it receives both speech and gesture commands. The aim was to investigate the usage potential of Kinect device where classical interaction techniques were not suitable and whether it could enable a more natural interaction by gestures and voice commands. This system showed that Kinect could be used as an alternative where traditional input methods of mouse or keyboard were not suitable such as medical environments or in some educational settings. As a future work, it was planned to evaluate the usability of the system with real users. In addition it was also planned to update the system to support different 3D images that was created by 3D drawing software and extend its use in other areas.[4]

In this paper, present two studies of people typing with gaze+dwell and gaze+click inputs in VR. In study 1, the typing keyboard was flat and within-view; in study 2, it was larger-than-view but curved. Both studies included a stationary and a dynamic motion conditions in the user's field of view. There are four main options used for text entry in VR: 1) removing the headset and using a physical keyboard to enter text, 2) pointing with an external controller to activate the keys on an on-screen virtual keyboard , 3) speech-to text-conversion, and 4) pointing by head motion. Taking the headset off and using a physical keyboard to enter text breaks the experience of immersiveness and the user may need to re-adjust the fit of the headset on their face. findings suggest that 1) gaze typing in VR is viable but constrained, 2) the users perform best when the entire keyboard is within-view; the larger-than-view keyboard induces physical strain due to increased head movements, 3) motion in the field of view impacts the user's performance: users perform better while stationary than when in motion, and 4)gaze+click is better than dwell only interaction.[5]

Mostly modern computer program designed with interaction models with pointing device and keyboard input. Keystroke shortcuts are hard to remember and some tasks cannot be easily accomplished using keystroke shortcuts, so pointing device is required. The approach used in this paper is intuitive interaction paradigm of graphical user interface (GUI) while eliminating subtask that make system slow, it is achieved by the Multi Model Interaction Concept for Efficient Input(M2ice Input) with constant gaze tracking with keyboard command. Most modern interaction model have two input devices and switching between two input devices cause stress and performance loss, so user request to reduce number of interactions. The Multi Modal interaction concept for efficient input based on combination of keystroke shortcuts as well as eye tracker as input device for detecting gaze direction.[6]

An immersive system provides unique experience in human-computer interaction by presenting user a virtual environment vividly. Existing system mostly depend on the visualization techniques for immersive experience, so the interaction between user and system is often limited which require the user to actively operate equipment. As a solution of this, Cognitive Immersive Room (CIR) has been introduced that supports multi model interaction without needing to use extra interactive equipment. By keeping track of spatial and temporal context, CIR provides foundation of higher level cognitive tasks like emotion, understanding, social behavior analysis, reasoning etc. In CIR, first, gesture recognition and face analysis techniques with human-computer interaction context have been developed. Second, an integrating system that allows fusion multi-model input to support natural interaction has been developed. Last, the flexibility of system in supporting different use cases has been checked.Besides using human scale multi- media environment, gesture, face and speech recognition techniques are also enabling for multi- model interaction between users and systems.[7]

Augmented reality has changed the way that everyone interacts with machines and with each other. Multiple interactions approaches are needed in AR application to enhance its user experience. Mostly two input modalities which is touch screen and inertial measurement unit(IMU) are used. In this paper two input modalities of camera and microphone to gesture and speech-based interaction have been mapped. The gesture interaction modality is enabled by Leap motion controller, and the speech interaction modality is enabled by Google cloud speech. One experiment had been conducted in which eighteen subjects has been participated, each subject performed certain tasks using speech command, gesture command or mixture of both, after that elapsed time and accuracy has been measured. As a result of this experiment certain statements have been made like 1) speech outperforms gesture in terms of accuracy 2) speech and gesture have about same elapsed time. 3) Gesture has relatively low accuracy. [8]

Human-robot interaction (HRI) is quite interesting research area nowadays because of growing intrusion of social robot in everyday life. Humans mostly communicate with the audio-visual information, so gesture and speech recognitions are highlighted HRI research areas. Child-robot interaction (CRI) is a part of HRI which is an interdisciplinary research area. In this paper they proposed a Multi3: (Multi- model, Multi-robot, Multi-sensory) system for robot perception mainly developed for CRI. Multi-model include action, speech, and gesture control while multi-sensory focus on different sensors like kinect camera, microphone arrays etc. During experiment child can interact with social robot naturally using body action, gestures and speech, then system employs multiple kinects using visual sensors and microphone to capture child's activity. After that Multi3 system has been evaluated, by using a questionnaire that was filled by children after their interaction with system. As a result the majority of children enjoyed interacting with robots and likes to continue playing with them. [9]

In recent years the demand for drones for civilian and non-civilian applications has been increased. So, to fully integrate them with society, it is little difficult to design safe and intuitive ways to interact with this aerial system. Aerial vehicles play fundamental roles for several applications. In this paper, a Graphical User Interface (GUI) and several Natural User Interface (NUI) methods are studied and implemented with the use of computer vision techniques. The two main mediums to implement reliable human-drone interactions (HDI) are voice and gesture based NUIs. In order to test and evaluate performance of NUIs, a real time experiments were conducted. These tests consists of individually testing proposed interaction methods in controlled indoor environments without aid of Motion Capture System. Along with that a multi modal interaction scenario is also tested by combining interfaces and permitting the user decide which interaction to use at given time. Visual body interaction, Visual Marker interaction, Hand gesture interaction, Speech command interactions, Multi modal interaction and NUI Diffusion is used for experiment. The result of experiment shows that use of NUIs for HDI are feasible options when operator and drone have higher level communication.[10]

Table 1: Comparative Table:

Sr. No	Paper Name	Year of Public	Technique name	Methodology	Advantage
1	Multimodal Human Computer Interaction: A Survey	2005	Large-Scale Body Movements, Gesture Recognition, Gaze Detection, Facial Expression Recognition, Emotion in Audio	1.Preprocessing The data. 2. Feature Extraction. 3.Apply different technique one by one. 4. Evaluation and Comparison of result.	we have discussed techniques applied in a wide variety of application scenarios, including video conferencing and remote collaboration, intelligent homes, and driver monitoring
2	An Intelligent Multimodal Mixed Reality Real-Time Strategy Game	2015	Mixed Reality	1.Leave the traditional turn-based style in favor of an increased simultaneous real-time interaction. 2.based on intelligent virtual agents. 3.controlled via a multimodal speech and gesture interface.	Such interfaces provide interaction alternatives given the dynamic nature of many applications in MR as well as in Virtual and Augmented Reality.
3	Emotion Recognition from Multi-Modal Information	2011	Facial Action, Coding System	1. Preprocessing the data. 2. Feature Extraction. 3.Apply different technique one by one. 4. Evaluation and Comparison of result.	This paper presents a fast, robust and accurate method for Emotion recognition
4	Multimodal Natural Interaction For 3d Images	2013	ANN	1.Image Capture 2.Image Processing 3.Feature Extraction 4.Classification 5.Evaluation and Comparison	There is a significant increase in the number of gestures and also now the gestures are more intuitive and user friendly. The accuracy of combined multimodal gesture recognition system is increased in comparison to the case when hand and head gestures were used
5	Gaze Typing in Virtual Reality: Impact of Keyboard Design, Selection Method, and Motion	2018	on-screen virtual keyboard, speech-to text-conversion, pointing by head motion	1.gaze typing in VR is viable but constrained. 2the users perform best when the entire keyboard is within-view; the larger-than-view keyboard induces physical strain due to increased head movements. 3.motion in the field of view	Users perform better when the entire keyboard is within-view. Motion in the user's field of view negatively impacts performance, induces strain, and some individuals may experience motion sickness. Though gaze+dwell based

				impacts the user's performance: users perform better while stationary than when in motion. 4.gaze+click is better than dwell only interaction	selection feels natural and easy, gaze+click is the most preferred way of interaction.
6	A Multi Modal interaction paradigm combining gaze tracking and keyboard.	2017	Graphical user interface, Keyboard and eye tracking	Uses combination of a handful of keystroke shortcuts as well as an eye tracker as input device for detecting gaze direction	The Multi Modal interaction concept for efficient input based on combination of keystroke shortcuts as well as eye tracker as input device for detecting gaze direction.
7	An Immersive System with Multi-modal Human-computer Interaction	2018	Cognitive Immersive Room	1.first, gesture recognition and face analysis techniques with human-computer interaction context have been developed. 2.an integrating system that allows fusion multi-model input to support natural interaction has been developed. 3.the flexibility of system in supporting different use cases has been checked.Besides using human scale multi- media environment, gesture, face and speech recognition techniques are also enabling for multi-model interaction between users and systems.	By keeping track of spatial and temporal context, CIR provides foundation of higher level cognitive tasks like emotion, understanding, social behavior analysis, reasoning etc.
8	Multimodal Interaction in Augmented Reality	2017	Gesture-based Interaction, Speech based Interaction, Mixture of both(gesture & hand)	1.Eighteen subjects participated in the experiment. 2.Each sub-ject was asked to perform four tasks 3 times. 3.In the first round, only gesture commands were allowed. 4.In the second round, only speech commands were allowed. 5.In the third round, a subject was free to use a	In addition to com-mon interaction modalities, we added two modalities, gesture and speech, to offer more natural communication between the user and the virtual character.

				mixture of gesture and speech commands. 6.Then result will be declared	
9	Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots	2018	Multiple kinetic based system	Audio visual input from kinetic sensors is fed into speech, gesture and action recognition modules to address challenging nature of child- robot interaction	It allow all type of user to work more intuitive and more effective because it eliminates the need to switch input devices during interaction
10	Natural User Interfaces for Human-Drone Multi-Modal Interaction	2016	Natural User Interfaces (NUI), Graphical user interface(GUI), Computer vision	It include speech, body position, hand gesture and visual marker interactions used to directly command tasks to the drone, it is based on leap motion controller	Drones evolved from touch to touch-less, by using speech, hand gestures, body position or visual markers, by adopting more affordable sensors, like the leap motion and small on-board monocular cameras.

IV FUTURE DIRECTION

Future work in this line of research will target implementing the different interaction techniques on Human- Multi-Drone Interaction. Here, users will have the ability to choose from different ways of interacting with multiple drones simultaneously.

V CONCLUSION

We have highlighted major approaches for multimodal human-computer interaction. We discussed gaze tracking, immersive system using multimodal, and multimodal in augmented reality, multimodal child interaction, natural interface for human drone, Multimodal Mixed Reality Real-Time Strategy Game, emotion recognition, natural interaction with 3D image, impact of keyboard design.

REFERENCES

[1] Jaimes A., Sebe N. (2005) Multimodal Human Computer Interaction: A Survey. In: Sebe N., Lew M., Huang T.S. (eds) Computer Vision in Human-Computer Interaction. HCI 2005. Lecture Notes in Computer Science, vol 3766. Springer, Berlin, Heidelberg

[2] Sascha Link, Berit Barkschat, Chris Zimmerer, Martin Fischbach, Dennis Wiebusch, Jean-Luc Lugrin, Marc Erich Latoschik, “An Intelligent Multimodal Mixed Reality Real-Time Strategy Game” IEEE Virtual Reality Conference 2016 19–23 March, Greenville, SC, USA.

[3] Chung-Hsien Wu, Jen-Chun Lin, Wen-Li Wei and Kuan-Chun Cheng,” Emotion Recognition from Multi-Modal Information”, IEEE 2015

[4] Fatih Ergüner, Pınar Onay Durdu, “MULTIMODAL NATURAL INTERACTION FOR 3D IMAGE”, IEEE 2016

[5] Vijay Rajanna, John Paulin Hansen,” Gaze Typing in Virtual Reality: Impact of Keyboard Design, Selection Method, and Motion”, ETRA ’18, June 14–17, 2018, Warsaw, Poland c 2018 Association for Computing Machinery.

[6] Luisa Brinkschulte, Robert Mertens, Leon Stapper, Sebastian Pospiech, Lars Knipping,” A Multi Modal interaction paradigm combining gaze tracking and keyboard.” 2017 IEEE International Symposium on Multimedia.

[7] Rui Zhao1, Kang Wang, Rahul Divekar, Robert Rouhani, Hui Su, Qiang Ji1, “An Immersive System with Multi-modal Human-computer Interaction”, 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition.

[8] Zhaorui Chen, Jinzhu Li, Yifan Hua, Rui Shen, Anup Basu,” Multimodal Interaction in Augmented Reality”, 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

[9] Antigoni Tsiami, Petros Koutras, Niki Efthymiou Panagiotis Paraskevas Filntisis,Gerasimos Potamianos, Petros Maragos,” Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots,2018 IEEE International conference on Robotics and Automation(ICRA) May 21-25, 2018, Brisbane, Australia

[10] Ram´on A. Su´arez Fern´andez1, Jose Luis Sanchez-Lopez1, Carlos Sampedro1,Hriday Bavle1, Martin Molina2, and Pascual Campoy1,” Natural User Interfaces for Human-Drone Multi-Modal Interaction”, 2016 International Conference on Unmanned Aircraft Systems (ICUAS) June 7-10, 2016. Arlington, VA USA.