Harnessing Power of Multimodal Interaction, their Challenges and Future Prospect – A Review

Dr. I Jeena Jacob

GITAM University, India.

E-mail: jeni.neha@gmail.com

Abstract

Multimodal interaction refers to the combination of smart speakers and displays. It gives users the option to engage with various input and output modalities. When interacting with other individuals, humans use more nonverbal cues compared to verbal cues. They communicate with each other using a variety of modalities, including gestures, eye contact, and facial expressions. This type of communication is known as multimodal interaction. A specific type of multimodal interaction called human-computer interaction (HCI) makes it easier for people to communicate with machines. Several studies employing the aforementioned numerous modalities will discover that machines could quickly interact with a person by disclosing their feelings or actions. The research presented here provides an in-depth overview of multimodal interaction, HCI, the difficulties and advancements encountered in this field, and its prospects for future technological improvement.

Keywords: Multimodal interaction, HCI, facial expression, Eye tracker, Gesture recognition, UCD, AR, VR.

1. Introduction

The multimodal human-computer interaction is defined as the interaction with the virtual and physical environment through spontaneous modes of communication. This suggests that multimodal contact connects humans with machines in both input and output, allowing for easier and morenatural communication. In particular, multimodal systems can provide a versatile, effective, and user-friendly environment that enables users to communicate with the system using input modalities like speech, handwriting, gestures, and gaze, and to obtain information

DOI: https://doi.org/10.36548/rrrj.2023.2.017

from it using output modalities like speech synthesis, intelligent graphics, and other modalities that are combined at the right time. Then, in order to enable their interpretation, a multimodal system must be able to identify the inputs from the various modalities and combine them in accordance with temporal and contextual limitations. This procedure is known as multimodal fusion. Human-computer interaction (HCI) research includes multimodal user interfaces [10].

1.1 Multimodal Input

The accessibility of multimodal input user interfaces is limited. A well-crafted multimodal application may be utilised by individuals with a wide range of disability types. Users who are blind or visually challenged utilise audio communication along with keypad input. Users with hearing impairments employ a combination of spoken input and visual modality. Other users will only employ the proper modalities as requested since someone is emotionally impaired (e.g., wearing gloves in a highly loud setting, driving, or attempting to enter a credit card number in a public area). This multiple input modalities could increase the usability when the advantage of one method outweighs the disadvantage of the other.

1.2 Multimodal Output

Redundancy and synergy are two of the multimodal output system's suggested advantages. Multimodal output is now employed primarily to facilitate attention management in data-rich environments where operators must meet high demands on their visual attention, as well as to improve the connection between communication medium and content. The late 1950s saw the earliest applications of touch as a communication tool. It's a potential and distinctive means of communication. The sense of touch is proximal—it feels items that are in contact with the body—and bidirectional—it supports both perception and action on the environment, in comparison to vision and hearing, the two classical senses used in HCI.

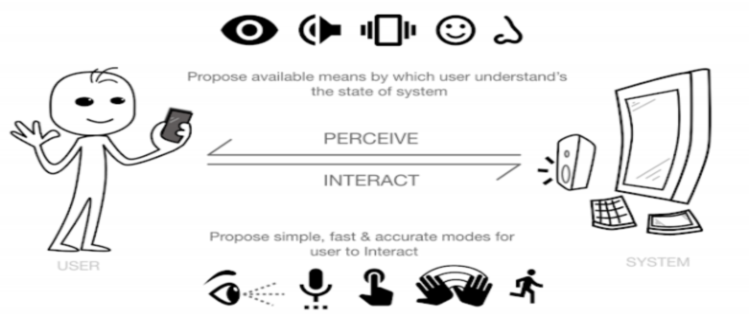
Propose available means by which user understand's the state of system
PERCEIVE
INTERACT
USER
SYSTEM
Propose simple, fast & accurate modes for user to Interact

Figure 1. Multimodal Interaction between Human and Machine [11]

2. Literature Survey

In this section, will be explaining the related works on the field of multimodal interaction with different technologies: -

Matthew Turk et. al. [1] reviews of multimodal interaction in the field. The primary topics covered in this research study are those related to multimodal interaction. Researchers talked about the lifecycle of multimodal interaction, including how it began and how it upgrades its capabilities in response to technological advancements. The potential and the challenges are described in depth. This research work was prepared specifically for the multimodal integration. The difficulties in the field of human-computer interaction (HCI) are thoroughly described.

In this research work [2] Multimodal interaction is one area where augmented reality finds use. Unimodal interaction is a common interaction strategy in augmented reality, allowing for single-user engagement. Many problems arise while employing unimodal in AR. Thus, this study demonstrates the utilisation of multimodal interaction in an augmented reality setting as well as the integrated elements incorporated into the multimodal interaction AR framework. The majority of the MMI AR framework's integrated components are only covered in this work. Future research projects in this area are also discussed.

In this research work [3], the design space for gesture and speech interaction with physically dispersed Internet of Things devices will be explored. The two comments—interaction and selection—that make up the users' IoT instructions are broken down in this suggested design area. The study work explains how the existing techniques can offer interaction through variations of speech and freehand motion for these two components. They used the design space to create a sensing platform proof of concept and show off the innovative IoT interactions.

Sreedha et. al. [4] suggested a problem of Visual Question Answering (VQA) involves multimodal interaction between two domains: Natural Language Processing and Computer Vision. The goal of the suggested effort is to create a system that might respond to an inquiry about a picture. The fundamental process involves CNN to extract the image features and RNN to extract the query feature. To obtain the most accurate results, multiple reasoning is applied to the retrieved characteristics. VQA can help blind individuals learn more about the world by serving as a visual assistant. This work provides a VQA assistant for blind people by applying a real-world dataset called the VizWiz dataset, which was generated by blind people.Carlos Busso et. al. [5] evaluates the benefits and drawbacks of systems that rely solely on audio or facial expression data. Additionally, it goes over feature-level integration and decision-level integration as two methods for combining these two modalities. Four emotions—sadness, rage, happiness, and neutral state—were categorised using an actress's database. The markers on her face allowed motion capture to record her precise facial movements while concurrently capturing her voice. The findings show that, given the emotions taken into consideration, the system based on facial expression performed better than the one based only on auditory input. The results also demonstrate how well the two modalities complement one another and how, when combined, the resilience and performance of the emotion identification system increase quantifiably.

3. Multimodal Interaction using Areas for Research

Humans use many perceptual modalities in both simultaneous and sequential ways while interacting with the world. This is known as multimodal interaction. For many years, multimodal human-computer interaction has worked to provide computers with comparable capabilities so that users may engage in more interesting, powerful, and natural-feeling interactions. Recent years have seen a tremendous advancement in non-desktop computing because of potent mobile devices and reasonably priced sensors. As a result, multimodal research that makes use of expression, contact, eyesight, and gesture is growing [7].

In today’s scenarios involving human-computer interaction, multimodal AI will be used to facilitate more intuitive and natural communication. Applications like voice assistants, which can comprehend spoken orders and react to them while also analysing visual clues from their surroundings, fall under a type of multimodal category [12].

This research has examined a number of aspects of VR's multimodality as well as cross-modal interactions between the various sense modalities and their attainable impacts. All of the advantages that multimodality in VR may provide are discussed in this study work. Numerous fields have taken advantage of these benefits to improve various VR applications, demonstrating that multimodality may in fact produce more lifelike and engaging VR experiences. The application situations, however, span several academic fields. The study primarily focuses on various multimodal applications to three fields where virtual reality has had a significant impact: entertainment, education and training, and the medical field [8].

4. Multimodal Interaction in HCI

The study of computer technology and its design with an emphasis on human–computer interfaces (HCI) is known as human–computer interaction. Researchers in human-computer interaction (HCI) study how people use computers and create solutions that let people use them in new ways. An apparatus that facilitates communication between an individual and a computer is referred to as a "Human-computer Interface (HCI)".The interface between computers and humans is essential to enabling the many ways in which humans and computers communicate. Other names for HCI include computer-human interaction (CHI), man-machine interaction (MMI), and human–machine interaction (HMI) [13].

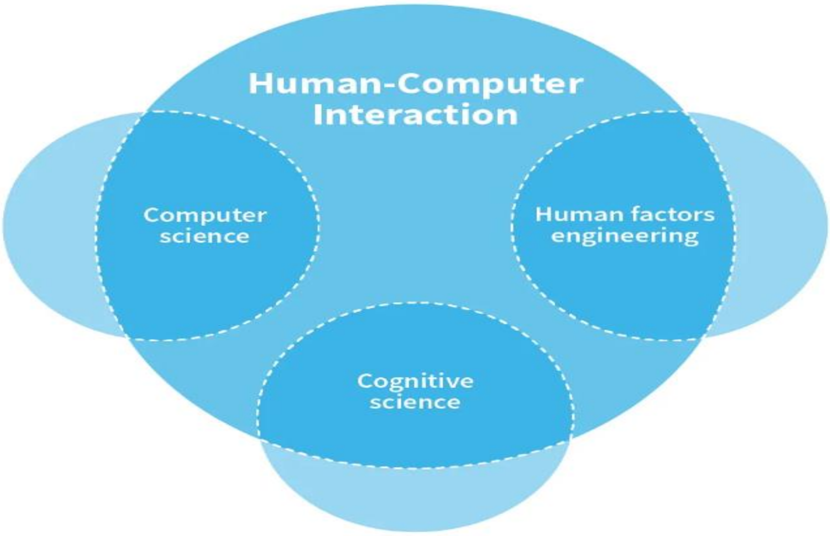

Figure 2. The Multidisciplinary Field of HCI [14]

A notion in Human-Computer Interaction (HCI) called "multimodal interaction" refers to the utilisation of several channels for human-computer contact or communication. Multimodal interaction expands on human-computer interaction (HCI) by combining many sensory modalities. HCI is primarily concerned with creating and enhancing human-computer interaction. A small number of multimodal interaction-related HCI aspects are:

4.1 Enhanced User Experience: Multimodal systems use speech, gestures, and touch, as well as other modalities of interaction, in an effort to deliver a more intuitive and natural user experience. This can improve the efficiency and human-like quality of computer interactions.

4.2 Adaptability: Multimodal systems are able to adjust to the choices and circumstances of the user.

4.3 Improved Accessibility: Interfaces that are multimodal can improve accessibility for people with different needs. For instance, people with certain impairments might find that using voice or gestures instead of conventional input techniques makes it simpler for them to communicate with a system.

4.4 Context Awareness: Multimodal systems are able to obtain a deeper understanding of the user's context by using data from several modalities.

5. Techniques in HCI

The multidisciplinary discipline of human-computer interaction (HCI) is concerned with the development and use of computer technology, with a special emphasis on human-computer interaction. In HCI, a variety of methods are used to enhance the accessibility, usability, and general user experience. Here are a few key HCI techniques:

5.1 User-Centered Design (UCD)

A group of procedures known as user-centered design (UCD) concentrate on placing consumers at the center of the design and development of products. To build products that are easy to use and accessible for users, UCD design teams incorporate people throughout the design phase using a range of research and design methodologies. For instance, a product team considers user needs, goals, and feedback when developing digital goods. User requirements and desires are prioritised, and each design choice is assessed in light of how well it will benefit the users. Your goods can have an emotional impact thanks to user-centered design [15].

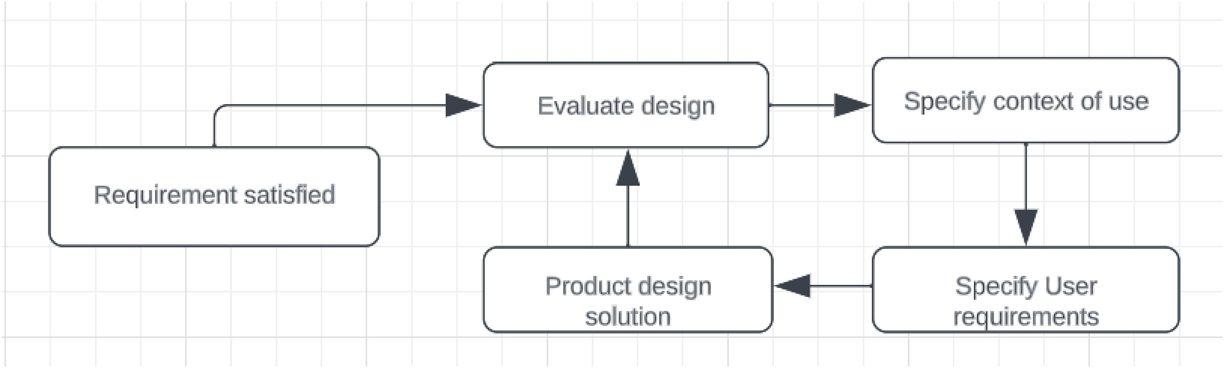

Figure 3. Phases of UCD

5.2 Gestural Recognition

With the use of sensors, gesture recognition technology is able to recognise and interpret hand gestures as commands. With the use of this feature, passengers and drivers may communicate with the car and often operate the entertainment system without hitting any of the buttons or screens. A camera is first aimed at a particular three-dimensional area inside the car in order to record frame-by-frame pictures of hand locations and actions for a gesture recognition system. Usually, this camera is installed in a roof module or another unobtrusive observation position. Even under low light conditions, the system uses infrared LEDs or lasers to illuminate the surrounding area and provide a crisp image [16].

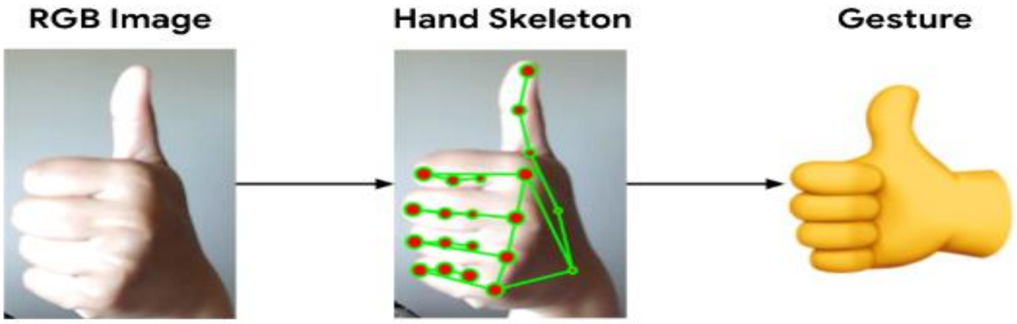

Figure 4. Hand Gesture Recognition [17]

5.3 User Research

Using methods like as surveys, usability tests, interviews, and other types of feedback, user research aims to comprehend user behaviours, requirements, and motivations. It is employed to ascertain if design solutions satisfy users' demands and to comprehend how users engage with things. By using experimental and observable research methodologies to inform product design, development, and refinement, this field of study seeks to improve the user experience (UX) of goods, services, or processes. Numerous goods, including websites, smartphones, medical gadgets, banking, government services, and many more, are improved by the application of user research. It is an essential component of user-centered design and an iterative technique that may be applied at any point in the product development process [18].

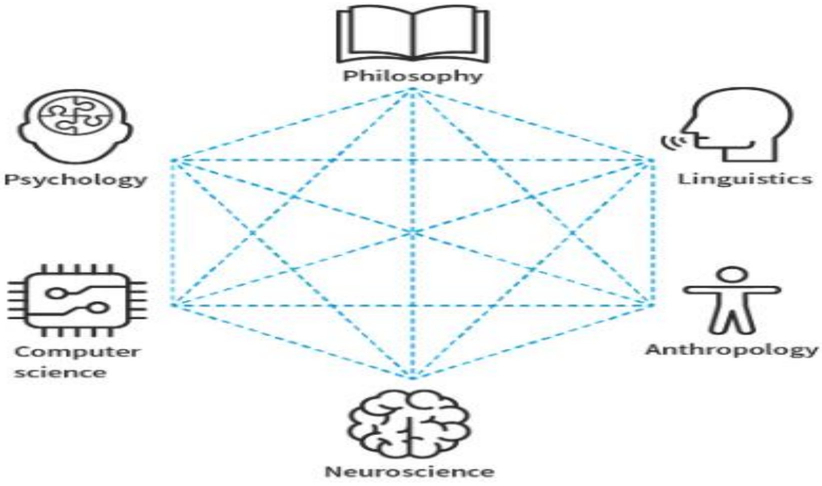

Figure 5. Different Field where User Research takes Place Regarding the HCI [14]

5.4 Virtual & Augmented reality in HCI

The linked areas of virtual and augmented reality with human-computer interaction (HCI) include technologies that enable interactions between people, computers, and the outside environment, hence enabling various uses of information technology that are advantageous to society as a whole. The goal of the area of human-computer interaction (HCI) is to build innovative hardware and software that can recognise and understand human behaviour and features in order to increase the efficacy and efficiency of human-computer interactions. Enhancements in HCI technology can result in more natural and effective methods for individuals to interact with a real or virtual environment, which can lead to improved experiences in virtual reality (VR) and augmented reality (AR).

The field of HCI encompasses a wide range of technical sub-specialties, such as machine learning for anticipating and satisfying user needs, virtual assistants, speech recognition, voice control, gesture recognition, behaviour recognition, behavioural analytics, mood/emotion recognition, tactile displays, haptics, biometric sensing, bioacoustic sensing, and biosignal recognition and processing. Wearable technology, such smart glasses, smart watches, and health monitors, may use these technologies [19].

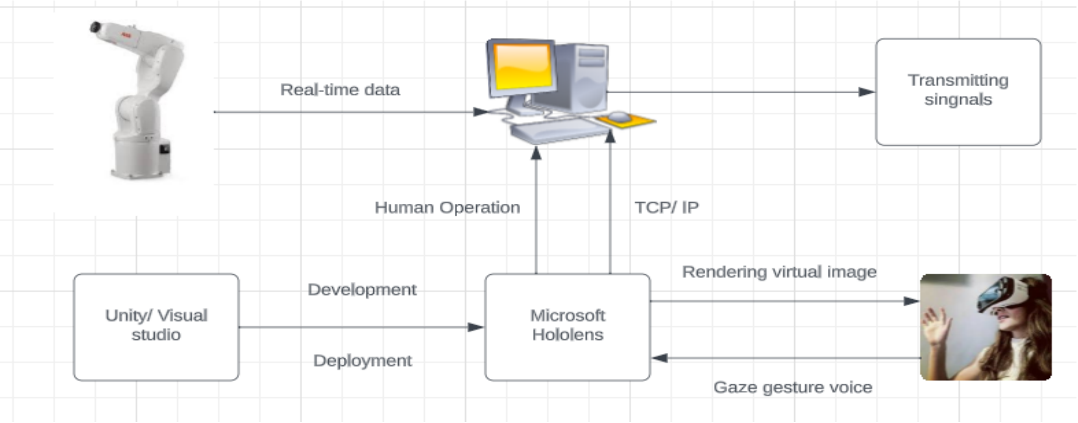

Figure 6. HCI with VR & AR

6. Challenges of Multimodal Interaction

Multimodal interaction has its own set of difficulties, even if it presents a potential path for improving human-computer connections. For multimodal systems to be successfully designed and implemented, several issues must be resolved. The following are some major obstacles:

6.1 Integration Complexity: It's difficult to effortlessly combine several mediums. Careful design and engineering are needed to create systems that can seamlessly combine voice, gestures, touch, and other modalities without becoming confusing or creating delays.

6.2 Environmental Variability: Depending on the user's surroundings, certain modalities may be more or less successful. For example, in loud settings, voice recognition could be less accurate. One major problem in system design is adaptability to changing environmental circumstances.

6.3 Cost and Resource Constraints: It could need a lot of hardware and processing power to implement reliable multimodal systems. It can be difficult to strike a balance between cost and performance, particularly in settings with limited resources.

6.4 Technological Limitations: Certain modalities, like voice processing or gesture recognition, may still have limits in terms of accuracy and dependability. Improvements in technology are required to increase these modalities' resilience and accuracy.

6.5 Privacy Concerns: Certain modalities give rise to privacy problems, such as gaze monitoring and facial recognition. In order to design systems that preserve user privacy while maintaining the necessary functionality, ethical and legal considerations must be carefully taken into account.

7. The Major Goals and the Progress of Multimodal Interactions

The basis of multimodal interface design has been established with the objective of enabling expressively meaningful, flexible, natural, and transparent ways of communicating with computers. It can boost performance in a range of activities, stimulate intellect, and lessen cognitive strain. User preference for multimodal interaction is higher than for graphical or unimodal keyboard-based interfaces. In a wide range of application fields, users strongly prefer to engage multimodally rather than unimodally. [6].

7.1 Flexibility: Multimodal interactions facilitate the different uses of input modalities. This involves deciding which modality to utilise to communicate different kinds of information, combining different input modes, or switching between modalities if circumstances change. For instance, even when voice is accessible, the user of an in-car application can often not be able to utilise manual or visual input. In the shifting conditions common to field use, a multimodal interface facilitates the selection and change of modalities as needed.

7.2 Efficiency: In the early days of multimodal system design, it was thought that the primary benefit of creating a multimodal interface would be efficiency advantages, which would result from processing input types concurrently. Similarly, compared to unimodal input, users' productivity increased when they combined voice and manual gestures to operate things in a multimodal manner. For instance, research demonstrates how multimodal voice and mouse input increase productivity when doing a sketching work.

7.3 Expressive power: According to recent research, interfaces with greater expressive capacity (such as multimodal or digital pens) can significantly boost cognition above and beyond what is possible with keyboard interfaces or comparable analog instruments. Pen input, for instance, is a rich creation of content modality that works with many kinds of representation, including verbal, symbolic, mathematical, and diagrammatic representations. While concentrating on a task, it may also be utilised to quickly and flexibly switch between representations.

7.4 Minimization of cognitive load: Multimodal interactions can significantly reduce users' cognitive burden and enhance performance from a cognitive perspective. One study found that during visual-spatial map tasks, those who communicated multimodally using voice and pen produced 36% fewer task-critical mistakes than those who interacted using speech alone.

8. Future Work

Exciting prospects for multimodal interaction lie ahead, and current research and development initiatives seek to push the envelope of what is feasible while addressing current issues. Future research in multimodal interaction might go in the following directions:

8.1 Enhanced Sensing Technologies: Improvements in sensor technology, such as more precise and reasonably priced cameras, microphones, and other sensing tools, can raise the accuracy and dependability of modalities like gaze tracking, voice processing, and gesture recognition.

8.2 Immersive Experiences in Virtual and Augmented Reality: In order to provide immersive and natural user experiences, multimodal interaction will become increasingly important as virtual and augmented reality technologies advance. The incorporation of voice commands, haptic feedback, and gestures can enhance the realism and interaction of virtual worlds.

8.3 Emotion Recognition: Computers may be able to comprehend and react to users' emotional states more effectively if emotion detection is included in multimodal systems. This may result in interfaces that are more adaptable and sympathetic.

8.4 Human-Robot Collaboration: In the future, multimodal interaction will be essential to human-robot cooperation. Creating interactions between people and robots that are both natural and efficient will be a major need as robots become increasingly embedded in daily life.

8.5 Explainable AI in Multimodal Systems: Transparency and explainability are becoming increasingly important as AI systems get more complicated. Future research might focus on creating techniques that improve users' ability to understand how multimodal AI systems make decisions.

8.6 Education and Training Applications: In training and educational environments, multimodal interaction may be used to provide more dynamic and captivating learning opportunities. Future research endeavors might investigate the creation of teaching resources that integrate many modalities.

9. Conclusion

This research work provides a quick explanation of multimodal interaction and the primary technology used in human-computer interaction (HCI). This article describes the multimodal interaction approach, the difficulties encountered in each field, and future studies and advancements that researchers desire to research. The main ideas and methods of HCI are thoroughly discussed. As technology advances, the industry keeps developing along with it, helping to create digital experiences that are more inclusive and user-friendly.

References

[1] Turk, Matthew. "Multimodal interaction: A review." Pattern recognition letters 36 (2014): 189-195.

[2] Nizam, SS Muhammad, Rimaniza Zainal Abidin, Nurhazarifah Che Hashim, Meng Chun Lam, Haslina Arshad, and N. A. A. Majid. "A review of multimodal interaction technique in augmented reality environment." Int. J. Adv. Sci. Eng. Inf. Technol 8, no. 4-2 (2018): 1460.

[3] Kang, Runchang, Anhong Guo, Gierad Laput, Yang Li, and Xiang'Anthony Chen. "Minuet: Multimodal interaction with an internet of things." In Symposium on spatial user interaction, pp. 1-10. 2019.

[4] Sreedha, B., and Prashant R. Nair. "Multimodal visual question answering using vizwiz data; a visual assistant for the blind." In International Conference on Electrical and Electronics Engineering, pp. 365-372. Singapore: Springer Nature Singapore, 2022.

[5] Busso, Carlos, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. "Analysis of emotion recognition using facial expressions, speech and multimodal information." In Proceedings of the 6th international conference on Multimodal interfaces, pp. 205-211. 2004.

[6] Oviatt, Sharon, and Philip R. Cohen. "Aims and Advantages of Multimodal Interfaces." In The Paradigm Shift to Multimodality in Contemporary Computer Interfaces, pp. 17-25. Cham: Springer International Publishing, 2015.

[7] Turk, Matthew. "Multimodal interaction: A review." Pattern recognition letters 36 (2014): 189-195.

[8] Martin, Daniel, Sandra Malpica, Diego Gutierrez, Belen Masia, and Ana Serrano. "Multimodality in VR: A survey." ACM Computing Surveys (CSUR) 54, no. 10s (2022): 1-36.

[9] Luan, Fengkai, and Xing Pan. "Human-machine integration interactive system based on mixed reality technology." In Journal of Physics: Conference Series, vol. 1549, no. 2, p. 022128. IOP Publishing, 2020.

[10] Multimodal_interaction_- https://en.wikipedia.org/wiki/Multimodal_interaction#:~:text=Multimodal%20human%2Dcomputer%20interaction%20refers,in%20both%20input%20and%20output.

[11] https://www.uxness.in/2020/04/getting-closer-to-multimodal-interaction.html

[12] Multimodal in AI - https://www.techopedia.com/definition/multimodal-ai-multimodal-artificial-intelligence#:~:text=Human%2Dcomputer%20interaction%3A%20Multimodal%20AI,visual%20cues%20from%20the%20environment.

[13] HCI - https://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction

[14] https://www.interaction-design.org/literature/topics/human-computer-interaction

[15] USD -_https://www.geeksforgeeks.org/introduction-to-ucd-user-centered-design/

[16] Gesture recognition- https://www.aptiv.com/en/insights/article/what-is-gesture-recognition#:~:text=A%20gesture%20recognition%20system%20starts,is%20unlikely%20to%20be%20obstructed.

[17] https://www.arxiv-vanity.com/papers/2111.00038/

[18] User_Research- https://en.wikipedia.org/wiki/User_research#:~:text=User%20research%20focuses%20on%20understanding,design%20solutions%20meet%20their%20needs.

[19] HCI with AR & VR - https://www.sbir.gov/node/1189909