IntelliHolo : 3D Modelled Artificial Intelligence with Mixed Reality; A Review

Pratham Shelke[1], Darshan Shah[2], Shivani Sakpal[3], Sanjana Joeel[4], Sagar Shinde[5], Anand Dhawale[6]

[1-6]*Department of Computer Engineering, M.E.S. College of Engineering (Wadia), Pune, Maharashtra, India*

---***---

Abstract - *Mixed Reality is a blend of the physical world and digital world, capable of bringing around changes which can help us better understand the world around us. Mixed Reality has already been proven capable in many business models. With this thought we are trying to integrate Artificial Intelligence into Mixed Reality, creating a human form in Mixed Reality, capable of understanding the world around it more like us, interacting with the world, more like us. This review focuses more on the building blocks of the 3D modelled Artificial Intelligence with Mixed Reality. The literature survey in this paper further supports the idea. Today every major commercial project uses Artificial Intelligence to either help implement business analysis or to better the user experience. With the merge of Mixed reality and Artificial Intelligence, we can go a step ahead of traditional voice recognizing intelligences, and create a totally new user experience. This could be a base on how robots can be made with more efficiency.*

Key Words: Mixed Reality (MR), Artificial Intelligence (AI), Robots, Three Dimensional (3D).

1. INTRODUCTION

1.1 Background and Motivation

The ability of a computer program or a machine to think and learn is known as Artificial Intelligence. It can be from learning the environment around itself to understanding user commands to even identify emotions [1]. Through the course of time Artificial intelligence has picked up a pace in its growth and improvement. This being possible due to standardization [2], the developers believing in the new technology, and trying to make something new out of it. Developers have developed not only chat boxes where a machine answers user's queries but also have developed mechanical robots which imitate human beings. Today we are all surrounded by Artificial Intelligence like Alexa, Google Assistant, Siri, Cortona, etc. What these do is understand a user command [3] and try searching the web for results and just dictate those results to the user. But when we are talking about Artificial Intelligence, how fair is it that we can only hear them and not see them, our future would not be like that. We see a near future where Artificial Intelligence can not only search the web for results but also be able to develop a thinking of its own and also understand the user not only logically but emotionally, like understanding sarcasm. First step towards achieving these goals would be to let the user interact with the Intelligence visually rather than just a voice. Mixed reality being the perfect ship for the Artificial Intelligence to be more visual and interactive, we are trying to build a 3D model of a human being [4] using Unity which would be our Artificial Intelligence in Mixed Reality, which can understand the environment around itself and can even interact with the things around it, say that there is a chair around it then the Intelligence should understand the purpose and use of the chair and should be able to perform actions accordingly, in this case seat in the chair. Artificial Intelligence being visual can play the role of an assistant, teacher and many more. It can reflect emotions and really connect with users, it can entertain the user, it can Play with users, and can perform many more functionalities. This forms the root of our motivation for this project.

1.2 HoloLens

Our primary hardware required for the system is Microsoft's HoloLens. It is a virtual reality (VR) headset with transparent lenses for a mixed reality experience. To be programmed using C sharp, with the support of unity, Maya and Blender, the models would be built and will be deployed in HoloLens to be rendered in Mixed Reality. The headset provides many functionalities like tracking motion, gesture recognition, etc.

2. LITERATURE S URVEY

[1] describes methods on how the system can understand the context of the speech, recognise the gender, whether the user is a male or female, and most importantly the user's emotion, considering many factors of a voice like frequency, amplitude, pitch, etc. The system in [1] is able to recognize the neutral state and six emotions namely-anger, boredom, disgust, fear, happiness, and sadness. The system proposed has further two subsystems (i) Gender Recognition and (ii) Emotion Recognition. Found from a survey that prior knowledge of gender increases performance of gender recognition. Pitch and amplitude of a voice signal is used to identify the gender of the user. Mean, variance, median, minimum, maximum and range of amplitude, pitch and speech energy is used to determine the emotion of the user. The system proposes use of open-source databases like, Reading-Leeds Database, Belfast Database, Berlin Emotional Speech (BES). A trained

dataset can further recognize emotions efficiently. Artificial Intelligence algorithms forms the basis on which the Mixed Reality model would enact and take decisions, so it becomes important to standardize the pillars of Artificial Intelligence on which our system would stand steadily. [2] suggests five pillars on Artificial Intelligence, the pillars being rationalizability, resilience, reproducibility, realism, and responsibility. [2] explains these pillars in depth, and also why we need them. On these pillars we can build a system which is stable and efficient. Fig 1 shows a diagrammatic view of pillars of Artificial Intelligence. Recent innovations in AI are mostly driven by ML techniques and revolve around the use of neural networks. Neural networks are said to be highly opaque which makes it difficult to interpret. With this in mind we need the modern AI systems to be rationalizable i.e. it should have the ability to be interpreted and explained. Recent research has proved that even an advanced AI system can be easily fooled, for this reason we need our AI to be resilient i.e. to attain high accuracy even when it faces adversarial attacks. Reproducibility of AI systems is the minimum necessary condition hence to overcome this AutoML has been introduced which is an attempt to develop algorithms that can automatically transfer and reuse learned knowledge over the systems. Realism of AI system refers to instilling the system with a higher understanding of emotional intelligence.

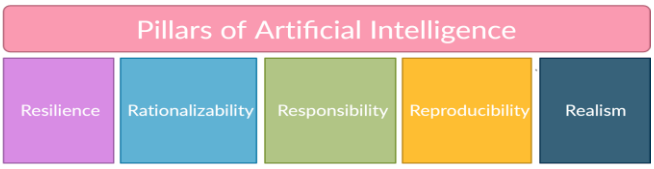

Fig -1: Pillars of Artificial Intelligence

[3] discusses the voice command recognition, of how to accurately and efficiently decode the user commands. The paper presents a system with an acoustic model (AM) trained with general speech dataset. The formation of grammar suggested by [3] is to have a set of every little mistake embedded in the command's grammar. Most important things to consider here is that every region has its own accent, like the Russians are sharper on r words, and then we have pronunciation differences for a single word, most importantly how a word is pronounced in America is different in the United Kingdom. Every aspect like these for a speech is covered in the paper, with the methods to overcome a faulty situation. The system also takes into view those words whose pronunciation are similar but means different in different ways, depending on the scenarios and also, they differ in spelling, words like set, said, sat, sait, sed will sound similar while speaking, another set of words would be pause, pose, pase, porse, pas, they would also sound similar, the challenge for the system here is to efficiently make out the words depending on the context that the speaker is speaking of. The system maintains a set of these similar sounding words, so that it can distinguish it. A set of grammar augmentation is shown in a Table 1 below.

Table -1: Candidate set for grammar augmentation

Command (c)	**Original grammar**	**Candidate set for grammar augmentation**
play music	play music	pla music, ply music, play music, …………
stop music	stop music	stap music, stup music, stop music, ……….
previous song	previous song	previs song, previous son, previous song, …….
next song	next song	nex song, lext song, nex son, …………

[4] discusses rendering 3D models of real humans into Mixed Reality. The paper discusses how to capture images of a human and then render a 3D model of him/her using efficient rendering algorithms, including the hardware setup. Using a 9-camera setup the images of the person are captured first, then feeding the images to the system, the 3D model is generated. It can even capture motions. The paper also presents several techniques to produce good quality and speed up the whole system. Pixel based algorithms are used to render human 3D models. The paper further discusses subtraction of the surrounding of the human while building the 3D model of the human. Fig.2 describes the overall process proposed in the paper.

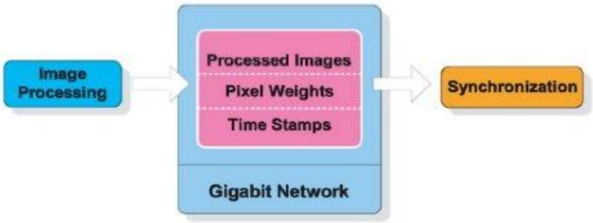

Fig -2: Data Transfer

Mixed Reality is a domain that is growing day by day, tackling the problems it faces, so when developing a project in this domain, it becomes important to keep a track of the technologies for the domain, the current trends in the market and most importantly the challenges faced, so we can be prepared beforehand. [5] proposes a survey on Mixed Reality with all possible current trends, and the challenges and in some cases how to tackle these challenges. The paper discusses overall about Mixed Reality as a domain, right all from hardware requirements to software integration, and also discusses algorithms that a Mixed Reality system would need for efficient processing of information and achieving it's ultimate goal, like rendering algorithms, object identification algorithms, object tracking, etc. Fig.3 describes the algorithmic requirement analysis.

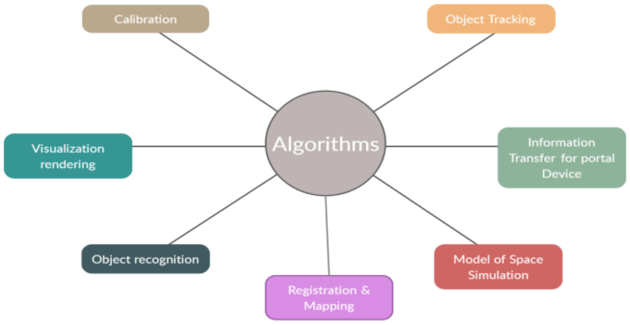

Fig -3: Algorithmic Requirement Analysis. Data Transfer

Mixed Reality system should be able to detect objects in the surrounding area. [6] discusses a problem of HoloLens lacking in providing high quality depth data, which restricts the use of HoloLens in some cases. The paper provides a solution on this problem, to use Intel RealSense or any RGBD camera mounted on HoloLens, connected to a stick PC, then to a power pack. The information is then feed to the processing PC, and then calculating accurate position of the 3D asset to be rendered on HoloLens. This increases the efficiency of a Mixed Reality system and further improves the rendering of the 3D assets, as it calculates the position of the 3D assets from the huge data that is received form the RBD camera. Fig.4 illustrates this method.

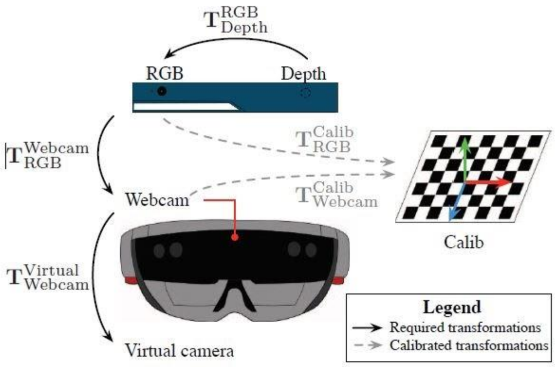

Fig -4: Mounting RGBD depth camera

[7] discusses Euclidean distance as a method for a system to understand numbers and letters, whether handwritten or computer generated. The image is processed to calculate pixel depth and accordingly treating the view as graph, it tries to calculate the Euclidean distance for every pixel, grouping it, then identifies the digit or letter on the sign. The method also can even be used in facial recognition.

When rendering an object in Mixed reality, we need to take care of the aspect of light direction, light intensity, and other factors related to light, so that the 3D model rendered looks realistic and provides a well-defined view. [8] provides a thorough discussion on how to render objects which are affected by the light. Say a 3D model of a glass object is to be rendered, then we need to consider aspects like light direction and light intensity, as these objects are special in a way which they reflect and refract incident rays in such a way that they cause a special effect known as caustic effect. On this basis the paper discusses how to select virtual point lights (VPLs) so as to determine the light source and indirectly the incident light direction. The paper proposes three steps to accurately render these special objects. One method handles more accurate reflections compared to simple cube maps by using impostors. Another method is able to calculate two refractions in real-time, and the third method uses small quads to create caustic Effects. Fig.5 shows caustic effect caused by transparent objects.

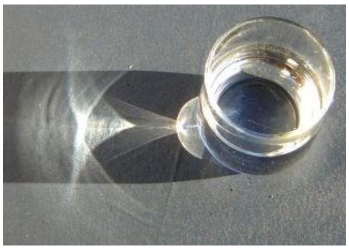

Fig -5: Caustic Effects

[9] provides a thorough approach of calculating the distance of the objects from the user by calculating the pixel density in the images received from the live feed of the surrounding. The paper intends in calculating the distance between cars in traffic and pedestrians and the car, so that accordingly the braking is applied. The image from the front camera of the car is taken is classified using a Convolutional Neural Network, namely YOLO (You Only Look Once) and after the searched object is detected, the distance is estimated counting the number of pixels in a bounding box which fits the detected object. The paper focuses on the safety of both the driver and the pedestrians.

[10] provides a revised discussion on how to differentiate between two or more users, the system should recognise who is speaking at a moment. Deep learning is suggested in this paper because of its strength in accuracy when trained with large amounts of data. Also provides a comparative study of traditional methods and deep learning-based methods. Siamese Neural Network (Siamese NN) proves to be best suited for their purpose. Fig.6 shows identification and verification process pictorially. Depending on the factors of a voice, like pitch, frequency and amplitude, the registered voices are differentiated.

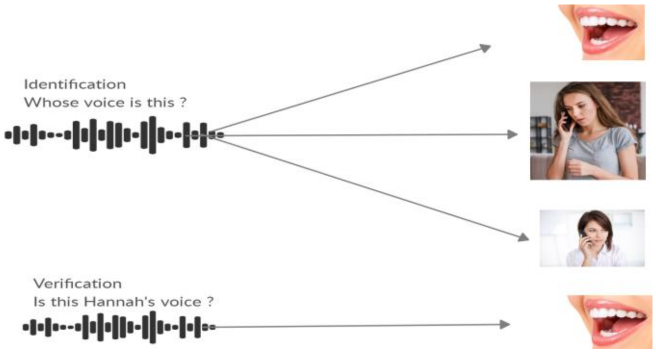

Fig -6: Identification and verification of voice.

3. PROPOSED SYSTEM

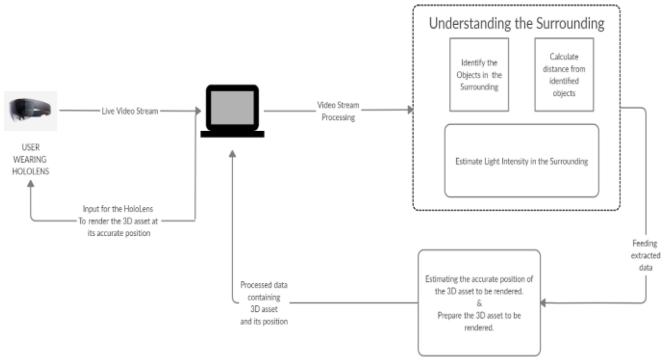
Understanding the Surrounding
Identify the Objects in the Surrounding
Calculate distance from identified objects
Estimate Light Intensity in the Surrounding
Live Video Stream
Video Stream Processing
USER WEARING HOLOLENS
Input for the HoloLens To render the 3D asset at its accurate position
Feeding extracted data
Processed data containing 3D asset and its position
Estimating the accurate position of the 3D asset to be rendered. & Prepare the 3D asset to be rendered.

Fig -7: Video Processing Architecture

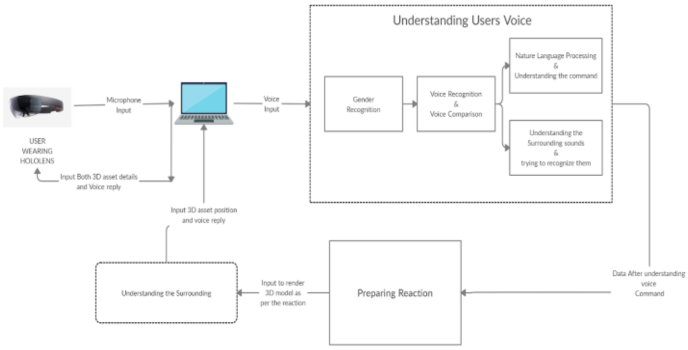
Understanding Users Voice
Microphone Input
Voice Input
Gender Recognition
Voice Recognition & Voice Comparison
Nature Language Processing & Understanding the command
Understanding the Surrounding sounds & trying to recognize them
USER WEARING HOLOLENS
Input Both 3D asset details and Voice reply
Input 3D asset position and voice reply
Understanding the Surrounding
Input to render 3D model as per the reaction
Preparing Reaction
Data After understanding voice Command

Fig -8: Voice Processing Architecture

1. Our project focuses on creating a user experience with a visually available Artificial Intelligence, for this to actually happen our system should be able to understand the user, emotionally.
2. Voice being the primary way of communication, the system must understand what the user is saying, whether he/she is asking a question, or whether he/she is ordering a task. The system should be robust enough to make out what the user intends to say.
3. Our primary goal is to make a 3D model of a human and then deploy it into Mixed Reality,
4. When we say that the system should be able to understand its surroundings then it should also be able to read signs around it, and understand them.
5. Understanding the surrounding covers another aspect, that is knowing how far are they from the subject, the subject being the 3D model in the Mixed Reality. This is important because when the subject explores the surrounding, we intend to add a functionality of interacting with the surrounding, for this the subject should know where and how far an object in surrounding is, to interact with it. Say there is a chair in the surrounding, and the subject is ordered to take a seat, then it has to walk to the chair, for which it has to know the distance to estimate the steps to be taken towards the chair and then have a seat.
6. The system should be able to differentiate between users, between the user's voice and noise and between noise and real surrounding sounds like horn, animals, birds, etc.

4. CONCLUSION

With the blend of Artificial Intelligence and Mixed Reality, we can look towards the world around us in a totally different way, where every little detail in our surroundings would be observed and processed. With our 3D modelled Artificial Intelligence people can now have a friend or assistant with them 24x7. An assistant that has the internet on its fingertip to get any information at any time. Not just information, but the intelligence can understand the user better and can be there even emotionally to cheer our user when he is sad. Our Proposed system targets to achieve at least 90% accuracy.

REFERENCES

[1] IGor Bisio, Alessandro Delfino, Fabio Lavagetto , Mario Marchese , AND Andrea Sciarrone on "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE Journal, pp. 1-14, 2013.

[2] Yew-Soon Ong and Abhishek Gupta on "AIR5 : Five Pillars of Artificial Intelligence Research", IEEE Journal, pp. 1-5, 2019.

[3] Yang Yang, Anusha Lalitha, Jinwon Lee, Chris Lott on "Automatic Grammar Augmentation for Robust Voice Command Recognition", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2019.

[4] Ta Huynh Duy Nguyen, Tran Cong Thien Qui, Ke Xu, Adrian David Cheok, Member, IEEE, Sze Lee Teo, ZhiYing Zhou, Asitha Mallawaarachchi, Shang Ping Lee, Wei Liu, Hui Siang Teo, Le Nam Thang, Yu Li, and Hirokazu Kato on "Real-Time 3D Human Capture System for Mixed-Reality Art and Entertainment", IEEE Journal, pp. 1-16, 2005.

[5] Somaiieh Rokhsaritalemi , Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi on "A Review on Mixed Reality :

Current Trends, Challenges & Prospects", MDPI Journal, pp. 1-26, 2020.

[6] Mathieu Garon, Pierre-Olivier Boulet, Jean-Philippe Doiron, Luc Beaulieu, Jean-Francois Lalonde on "Real-Time High Resolution 3D Data on the HoloLens", IEEE Conference Paper, pp.1-3, 2016.

[7] Liwei Wang, Yan Zhang, and Jufu Feng on "On the Euclidean Distance of Images", IEEE Journal, pp.1-6, 2005.

[8] Martin Knecht, Christoph Traxier on "Reflective and Refractive Objects for Mixed Reality", IEEE Journal, pp. 1-7, 2013.

[9] Gafencu Natanael, Cristian Zet, Cristian Foalu on "Estimating the distance to an object based on image processing ", IEEE Conference Paper, pp. 1-6, 2018.

[10] Nishtha H. Tandel, Harshadkumar B. Prajapati, Vipul K. Dabhi on "Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey", 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 1-7, 2020.

[11] [online] http://worldwidescience.org

[12] [online] http://export.arxiv.org/