Real world objects augmentation in virtual 3D environment RealSense SDK, Deep Learning and Game Engine

by

Shadman Shahriar Hasan Arko
20141039
Samiul Islam Shad
16301068
Md. Tamjid Hossain
16301217
Faysal Ahmed
17201072

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2020

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shadman

Shadman Shahriar Hasan Arko
20141039

Samiul

Samiul Islam Shad
16301068

Md. Tamjid Hossain

Md. Tamjid Hossain
16301217

Faysal

Faysal Ahmed
17201072

Approval

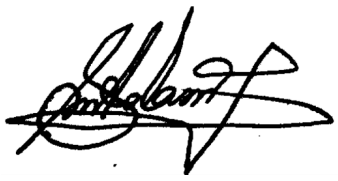
The thesis titled "Real world objects augmentation in virtual 3D environment using RealSense SDK, Deep Learning and Game Engine " submitted by

1. Shadman Shahriar Hasan Arko (20141039)
2. Samiul Islam Shad (16301068)
3. Md. Tamjid Hossain (16301217)
4. Faysal Ahmed (17201072)

Of Summer, 2010 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 05, 2020.

Examining Committee:

Supervisor:
(Member)



Md. Ashraful Alam, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

We propose and demonstrate an approach for real world objects augmentation in a virtual 3D environment using deep neural networks and game engine. The proposed system consists of three basic parts: acquisition, processing and visualization with VR. The processing unit comprises translation of 2D to 3D images of real world objects and augmentation of the real world object in a virtual environment with unity. Through deep learning and RealSense SDK, skeleton data is achieved which helps augment the real world 3D object into a virtual environment in unity. The proposed system enables real-time augmentation of the real world objects in a virtual 3D environment created in unity game engine. Therefore, the system allows users developing mixed reality applications for real life problem solving.

Keywords: VR; Machine Learning; Skeleton; Depth image; Prediction; 3D; Real Sense; Unity

Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our advisor Md. Golam Rabiul Alam sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$2D$ Two Dimensional

$3D$ Three Dimentional

AR Augmented Reality

$DOTS$ Data-Oriented Technology Stack

$DOTS$ Data-Oriented Technology Stack

ECS Entity Component System

GPS Global positioning System

ML Machine Learning

VR Virtual Reality

Chapter 1

Introduction

1.1 Motivation

The time we are living now would not sound so real if anyone told us about it in the past. It has been almost 7-8 months since we are staying at home without any having to go for work or studies. We cannot go out even if we want to, because of this deadly and contagious pandemic. All means of office and education has been taken online. Never before more than now we felt how important online meetings can be. This is happening on a mass level. However, even google meet and zoom meetings sometimes or for some people don't feel adequate to satisfy our needs. According to psychology, 'the more senses we use while encoding or learning, it stays in mind for longer and becomes easier to understand and focus'. Now is the time, where important meetings cannot be held in person. In this time, VR meetings can be a big achievement. According to our proposed method, any real-life object will be taken as an input. Then through deep learning and RealSense SDK algorithms, successful real-time mirroring of the object in front of the camera in a 3 dimensional model world will be possible.
The works we have seen in this area are VR games and applications. They use a specific video format which lacks a 3D environment and accurate 3D modeling. With our proposed model, a 3D model inside the 3D environment mirrors all the gesture and posture of the real-world model in real-time. This means being able to participate in virtual meetings where we can not only just see and hear each other like google meet but also being able to see each other's gesture and posture which will showcase the integrity and focus of each participant of the meeting. In out zoom classes what we have faced most is the lack of response. Teachers go around explaining topics with or without his video turned on and with the audio. However, he only has to rely or most times guess that students have understood because not many of the students respond. The teacher is always having the feeling of the confusion about whether he was able to deliver his lecture properly to the students. In this case, if our system is implemented, the whole system will ensure a lot of involvement of both the parties in the process. Even, during online exams, where the faculties are struggling to restrict cheating, the proposed model may come in handy as well. With the future improvements of the proposed model, it will make sure more involvement and will give the feeling of a the real-world examination scenario online.

1.2 Proposal

The proposed model requires a depth camera for the RGBD value of the real-world object where RGB stands for red, green, blue and D stands for depth respectively. This depth data associated with each pixel, gets sent to RealSense SDK, with its in-build hybrid hand and body tracking algorithm and with the help of deep learning the skeleton data of the rea-world object is generated. Then the skeleton data gets processed in Unity for visualization in VR. Another branch goes after the data acquisition to 2D to 3D translation which is the future development of the proposed model. It will use the RGBD data to convert 2D images to 3D using generative algorithm and in the end it will get processed in Unity for visualization in VR.

Chapter 2

Literature Review

For our research work, we had to study a bunch of earlier research paper related to Skeleton Depth image, Real time Texture and VR which helped us to find our way to accomplish our research work.
In another research paper, the main focus is using advanced anatomic human model to increase the strength of motion capturing. The motion capture in movies like 'Titanic' does a magnificent job at mirroring fake humans. While in games the highlight is on the strength of the rebuild system since in games there are a lot of human interactions. Sometimes while motion capturing important markers are not included or it mistakes one marker's trajected path with another's. The paper focuses on anatomic human modeling and its strength. The important goals are marker's accuracy, visibility and mobility of the skeleton. The authors combined marker rebuild in 3D with recovery of motion to avoid ambiguity from beforehand. When there is not enough surety of rebuilding from the prediction of the image trajectory, they have inclined towards the skeleton's anticipated location for the marker's identification and removal of the ambiguous 3D position. The few motion capture techniques it discussed are On line and Off line motion captures. For taking the input of the subject, the subject had to wear markers then the skeleton data is processed from that. The next step of the model is scaling the skeleton to subject's proportions while also ensuring subject's marker's position based on joints. They used gym motion which is the subsequent movements of the body those involve primary joints of the body. So in the marker rebuilding in 3D they focused on dividing the gym motions with one movement of one body part. Then the marker is rebuilt in 3D from 2D through stereo triangulation[2]. Then the 3D position of major joints in the frames in the sequence is projected with the help of the markers[37]. Then key joints 3D locations are achieved. After the skeleton gets initialized and scaled to subject's body proportion, complex motions are captured. The skeleton model is projected based on the observational data. Then the complex motions are captured before checking visibility and occlusion of the skeleton. After checking, 3D rebuild of markers in each frame is done by tracking the preceding frame into the current one. The result shows that in skeleton tracking 100% markers are reconstructed in each frame whereas in simple tracking most of the markers are not found in around 200 frames. Thus, the skeleton based motion tracking has far more strength and accuracy than the simple motion tracking[37].
In one of the studies it was proposed that, a method for estimating human full-body pose from the depth data gotten from a Time of Flight (ToF) camera. The method-

ology consisted of fitting anatomical landmarks into a skeleton humanoid model through robust detection using constrained inverse kinematics that is entirely built upon graph-based representation of the depth data which allows to measure geodesic distances between different parts of the body.The optical flow between subsequent intensity images provides motion information that is used to differentiate body parts when they overlap one another and thus provides a qualitative evaluation of the human body pose tracking method[7]. Another study was about generating monocular pose estimation framework through fusion of inertial and depth information. The model determines the body parts that are visible to the depth camera and tells reliable data modality resulting in estimation of simple and even complex humanoid poses. Moreover, depending on the body part visible pose parameters are identified with a generative tracker that uses a combination of inertial cues and optical cues. These optical and inertial cues are handled in different recovery schemes in order to retrieve pose databases while the preferential tracking is still in process. During the last combination step where results of both tracking methods mentioned above are used to obtain the final pose. Finally the tracker was assessed on a broad dataset which included inertial sensor data, adjusted depth pictures mocap system where it precisely caught poses even under stark impediment[10]. Another study was about generating monocular pose estimation framework through fusion of inertial and depth information. The model determines the body parts that are visible to the depth camera and tells reliable data modality resulting in estimation of simple and even complex humanoid poses. Moreover, depending on the body part visible pose parameters are identified with a generative tracker that uses a combination of inertial cues and optical cues. These optical and inertial cues are handled in different recovery schemes in order to retrieve pose databases while the preferential tracking is still in process. During the last combination step where results of both tracking methods mentioned above are used to obtain the final pose. Finally the tracker was assessed on a broad dataset which included inertial sensor data, adjusted depth pictures mocap system where it precisely caught poses even under stark impediment[12]. In another study, a model-based tracking of 3D pose and articulation of the human body formulated to minimize the difference between an observed 3D humanoid model and the model generated through observation of the hypothesized human 3D model. The observed humanoid model input comes from two RGBD sensors. The depth maps of the two RGBD sensors are combined to generate a volumetric representation of the human model. Although through standard techniques of implementing multiple conventional cameras and computing the visual shape of the human body figure, using two RGBD cameras provide less complexity and is cheaper[8].

As we have mentioned earlier that we are using real time texture in our research From the paper of Liang, Lin and Liu, Ce and Xu e.t.c the innovation is identified with a framework for productively textures surfaces from an input Sample, and specifically, to a System and technique for continuous Synthesis of high-calibre surfaces utilizing a fix based Sampling System intended to dodge potential component confounds across fix limits[3].

They also mentioned that A significant inspiration for surface Synthesis originates from the surface guide ping. Surface pictures, as a rule, originate from Scanned photos, and the accessible photos might be as well. Little to cover the whole article Surface. For instance, one ordinary Scheme utilizes a surface Blend approach that,

albeit dependent on a Markov Random Field (MRF) model of a given info surface, maintains a strategic distance from unequivocal likelihood work development and subsequent Testing from it. A related Scheme develops the previously mentioned Scheme by accommodating Verbatim duplicating and utilization of Small pieces, or on the other hand "patches" of the info Sample, instead of the utilization of individual pixels, for Synthesizing the yield surface[3].
From the paper of Method and apparatus for texture generation Yan, Johnson K. e.t.c mentioned that aeroplane pilot test programs give a choice to direct preparing to the utilization of aeroplane, the utilization of which may, on occasion, be awkward, uneconomical and additionally risky. Clearly, the legitimacy of such preparing is a circuitous connection to how much an aeroplane flight the test system mirrors the separate aeroplane including the inside of the cockpit, the reaction of the controls and, obviously, the view from the window[1].
They also mentioned that at that time current development, the favored strategy for surface age incorporates the underlying strides of getting surface regulation powers, sifting the forces to forestall associating, and putting away the separated powers in a two-dimensional surface tweak force table. The preferable technique for getting surface balance powers is from digitized genuine flying photos[1].
From an article we came to know that Image synthesis called as an old issue in computer and vision. The key difficulties are to catch the structure of complex classes of pictures in a concise, learnable model, and to discover effective calculations for learning such models and synthesizing new picture information. their methodology is like old style Generative Adversarial Networks (GANs), with the key contrast of not working on full pictures, yet neural patches from a similar picture. Likewise, they additionally, supplant the Sigmoid capacity and the twofold cross-entropy standards from by max-edge measures (Hinge misfortune). This maintains a strategic distance from the disappearing angle issue when learning. This is more hazardous for our situation than in Radford et all in view of less assorted variety in our preparation information[14].
In another research paper, one of their objectives is to examine the conjunction of immersive VR and unique representation, and the transaction between novel innovation and human recognition to produce understanding into both. VR and conceptual information perception have each autonomously been shown to improve science results, however to the best of information, there is no assessment of how vivid augmented experience can utilize unique portrayal of high-dimensional information on the side of logical examination. So straightforwardly the possible preferences of an immersive VR information show over the customary pictures, they have utilized vivid perception of Martian scene to give new capacities to planetary researchers, and telerobotic administrators. Their primary target was to explore whether the vivid condition furnished researchers with a more instinctive comprehension of the territory[11].
One of the studies describe that The basic programming and equipment are necessary for creating intelligent games to incorporate massively multiplayer web-based game structures, game engines and devices, streaming media, next-generation consoles, and remote and cell phones. Additionally they also mentioned that computer generated independence includes model-ing human and hierarchical conduct in arranged games, for instance, conveying the innovation from a game like The Sims for a genuine reason, for example, a preparation help for nursing. Engineers can

utilize this way to deal with the show and recreate medical clinic activities in-game structure, giving a immersive experience to the attendant student[4].
From a research we came to know that One of the embodiment's contain methods and systems to connect one another for communication in real-time within a virtual environment. The VR world hosted by the VR server that presents advertisement to a client inside the virtual environment where a reference is designated to the advertisement, one or more than one telecommunication server to form a bridge between the customer and advertiser for communication in real-time, a session border controller to interact with the packet switched network[35]. The virtual environment contains 3D object models and the advertisement to be showcased to the customer according to the customer's view point. Whereas, another embodiment says an object in virtual reality can sense and analyze the condition of its surroundings in order to showcase suitable advertisements to its customer avatars that may catch their attention. The term 'sense' refers to listening or observing conversations of nearby avatars to determine topics that might interest the avatars and present relevant advertisements[38].

Chapter 3

Fundamentals

3.1 Image Processing

3D image processing is the perception, handling, and investigation of 3D picture information through mathematical changes, sifting, picture division, and other morphological activities. With 3D image processing, it is conceivable to display structures of incredibly high unpredictability, for instance, anatomical structures in the human body, microstructures in a material example, or imperfection ridden modern parts. By making a precise output determined computerized model of the subject, testing issues can be tackled through basic investigation and recreation, for example, the structure of patient-explicit inserts or careful plans, enhancement of material structures for target properties, or non-damaging testing of high-esteem parts. Also there are few example of 3D image processing like handling of genuine 3D pictures, Appreciation of 3D items and 3D scenes, Creation and show of semi-3D pictures in 2D, Creation and show of 3D pictures utilizing optical and visual impacts, Stereology[5].

3.1.1 Characteristics of 3D image processing

1.Contrasted with human acknowledgement and comprehension of pictures: Humans doubtlessly, experience a cycle of joining 2D pictures, cross areas, to intellectually assemble a 3D structure. Carefully, we can't yet ensure that such a cycle is as a matter of fact happening, yet when seen from the external it appears as though that is surely the case. For computer, then again, moving from two measurements to three is essentially a standard augmentation to the preparing of the information held in a cluster, and isn't generally extraordinary. Computer can undoubtedly get to some random picture the component in either a few measurements, however people can't straightforwardly see the internals of a 3D picture. Besides, in light of the fact that genuine 3D pictures to incorporate all data about the inside structure of the subject, complex preparing identified with the recuperation of a 3D picture from a 2D source, for example, impediment or stereopsis is pointless. In such a manner, computer handling is indeed, even desirable over human vision[5].
2.Amounts of data: As contrasted and 2D pictures, 3D pictures speak to enormous measures of information. A single X-beam CT images, for instance, might be 512512512 bytes, while pictures from tissue test microscopy might be as extensive as 3000 2000 500 bytes. This effectively affects all parts of the picture handling

framework. As of now, information move among sensors and processors represents a noteworthy issue in CT gadgets[5].
3.The importance of display technology: The utilization of fitting presentation rules is in this manner critical, not just while showing the consequences of handling input picture records, yet in addition in specialists' assessment of middle of the road results. Preparing data contained inside a 3D picture for show onto a 2D screen is it could be said something contrary to the issue presented by PC vision. This innovation is presently inside the domain of PC illustrations or of data perception. It has a profound association with computer-generated reality and along these lines has a considerably more noteworthy essentialness in this time of sight and sound handling[5].
4.Properties of image geometry: In general the topological attributes of figures are very intricate. Therefore, the plan and assessment of related calculations, for instance, pivot (surface) diminishing, edge following, and so on. Turns out to be more troublesome.
5.Diversity of density values: The physical significance of thickness esteems isn't restricted to picture subjects and their chronicles, yet will differ significantly as indicated by space they are contained in and the estimation innovations utilized, in any event, when a similar subject is being dealt with[5].

3.2 Generative Algorithms

In our work field we are using 3d image processing in generative algorithm. 'Generative Algorithms' is a stage for research by structure and analysis. Through Generative Algorithms, it is expected to investigate such frameworks more top to bottom, with particular techniques, calculations and physical and computerized tests expected to understand a portion of the possibilities of such frameworks for additional improvement and use in configuration issues[6]. Mainly Generative model has a considerably much more complex task to perform. It needs to comprehend the conveyance from which the information is acquired and afterwards needs to utilize this comprehension to play out the assignment of grouping. Additionally, generative models have the ability to make information like the preparation information it got since it has taken in the circulation from which the information is given[22]. For instance in the event that I show a generative model a lot of canine and feline pictures now, the model ought to see totally what are the highlights that has a place with a specific class and how they can be utilized to create comparative pictures. Utilizing this data it can do numerous things. It can contrast the qualities with order the picture like how discriminative calculation can arrange a picture. It can create another picture which appears as though one of the class pictures it has been accommodated preparing. Progress in generative calculations is significant in light of the fact that people don't act simply like unadulterated discriminators, we have gigantic generative or creative abilities. In the event that we give certain traits, for example, blue vehicle on street we can immediately create an image of that in our psyche and we are taking a gander at giving this sort of knowledge to machines[22]. Definition of the Generative Algorithm is that it is a method of recounting to a tale about information; about the source of that information. Let's assume we have seen some information, at that point a generative strategy gives a potential clarification regarding how the information may have been produced. Probabilistic Inference is

(more often than not) the assignment of deciding the Cause given the perception. For instance, we have a mail (the perception) and we need to induce whether the mail is spam or not (the reason). There are fundamentally two ideal models for this undertaking.
1.Discriminative models legitimately model the P(Cause/Observation). Models like SVM and Logistic Regression do this.
2.Generative Models model P(Observation/Cause) and afterwards use Bayes hypothesis to PC P(Cause/Observation).

3.2.1 Generative Learning Algorithms

Generative methodologies attempt to assemble a model of the positives and a model of the negatives. You can think about a model as an "outline" for a class. A choice limit is framed where one model turns out to be almost certain. As these make models of each class they can be utilized for age. To make these models, a generative learning calculation learns the joint probability appropriation P(x, y)[17]. The joint probability can be written as:

$$P(x, y) = P(x|y).P(y) \tag{3.1}$$

Also, using Bayes' Rule we can write:

$$P(y|x) = P(x|y).P(y)/P(x) \tag{3.2}$$

Since, to anticipate a class name y, we are just inspired by the arg max, the denominator can be taken out from (ii). Consequently, to predict the mark y from the preparation model x, generative models assess:

$$f(x) = argmax_y P(y|x) = argmax_y P(x|y).P(y) \tag{3.3}$$

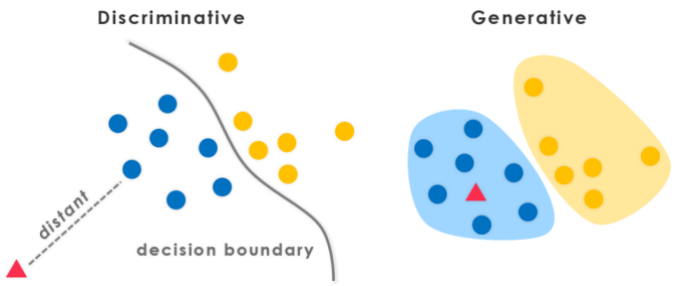


Figure 3.1: Generative Algorithms[17]

3.3 Machine Learning

Machine Learning is the investigation of Computer algorithms that improve naturally through understanding. It is viewed as a subset of Artificial Intelligence. Generally the main motive of the Machine Learning for the most part is to comprehend

the structure of information and fit that information into models that can be perceived and used by individuals. There is a little bit difference between traditional computer approaches and machine learning. In traditional computing algorithms are sets of expressly customized directions utilized by computer to ascertain or issue fathom. ML calculations rather consider computer to prepare on information data sources and utilize factual examination so as to yield esteems that fall inside a particular range. Along these lines, ML encourages computer in building models from test information so as to computerize dynamic cycle's dependent on information inputs[32].

3.3.1 Types of ML

1. Supervised Learning: Generally supervised learning is a learning where we instruct or train the machine utilizing information which is well marked that implies some information is now labeled with the right answer[30]. It assists with upgrading execution models with the assistance of experience. Manage to solve different sorts of real-world calculation issues. On the contrary, classifying large information can be tough task. Preparing for regulated adapting needs a great deal of calculation time. Thus, it requires a tons of time.

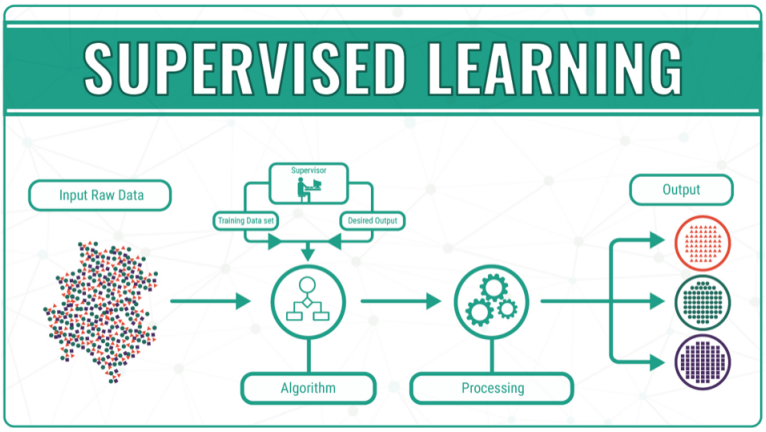


Figure 3.2: Supervised Learning[25]

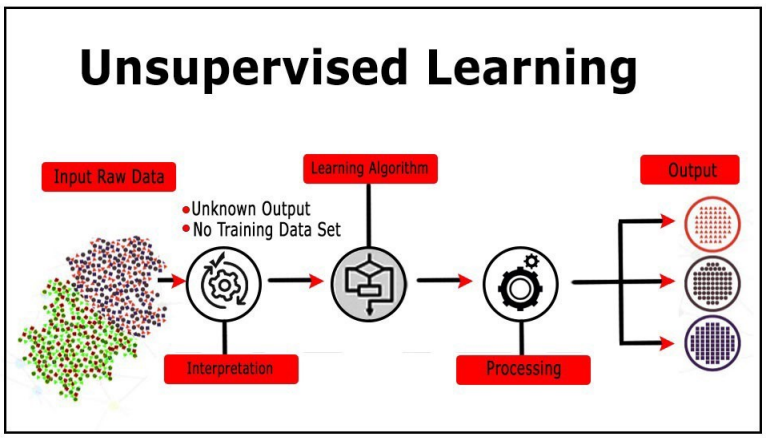


Figure 3.3: Unsupervised Learning[34]

2. Unsupervised learning: It is the preparation of machine utilizing data that is neither characterized nor named and permitting the calculation to follow up on that data without direction. It is utilized against information which isn't marked, computationally complicated and less accuracy[30]

3.4 Augmented Reality

Augmented reality (AR) is an encounter where designer improve portions of clients' physical world with computer created input. Planners make inputs going from sound to video, to illustrations to GPS overlays and then some in computerized content which reaction progressively to changes in the client's condition, ordinarily development. AR targets streamlining the client's life by bringing virtual data not exclusively to his prompt environmental factors, yet additionally to any aberrant perspective on this present reality condition, for example, live-video stream. In recent times, increasingly more globally eminent exploration organizations, colleges and ventures have put resources into the examination of AR, distributed bunches of papers and logical exploration results. These outcomes show the possibility and advancement of AR as human-computer cooperation innovation. With the improvement of the figuring intensity of computer programming and equipment, AR has continuously moved from the hypothetical exploration phase of the research Centre to the phase of mass and industry application, and as a scaffold between the advanced world and this present reality, it furnishes individuals with another approach to perceive and encounter the things around[21].

Figure 3.4: AR Game[16]

In addition, it has been listed as one of the top ten most promising technologies in the future by authoritative organizations such as the American Times Weekly[9].
Augmented reality overlays digital information on top of a camera-captured natural environment. Pokemon go is one of the best example of AR. It shows the actual AR.
To work, it needs the following components:

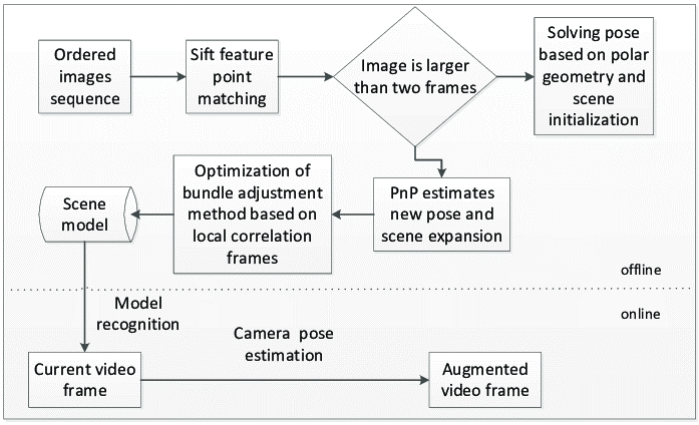


Figure 3.5: AR Workflow[19]

3.4.1 Methodology of AR

1. Depth-sensing camera: we need a camera to record visual data to add to a current article or spot. The said camera ought to be equipped for making sense of the subject's separation and edge from it.
2. Computer vision: As you utilize the camera, it takes pictures from the rest of the world for translation and referring to utilizing a ML. For instance, when you train a camera's emphasis on a container, it utilizes the pixels from that picture as a kind of perspective to perceive comparable looking objects. Whenever we snap a picture of another case, the algorithm would attempt to review this data to figure if the article is undoubtedly a crate. The ML algorithm additionally joins all pieces of data and adds an innovative touch to them to give the client a vivid and reasonable AR experience[33].
3. Output device: This refers to the display device where users can view the resulting image or video, such as a phone or computer monitor[33].

3.4.2 Types of AR

1. Marker based AR: Marker-based AR encounters require a static picture likewise alluded to as a trigger photograph that an individual can examine utilizing their smartphone through an increased reality application. The portable sweep will trigger the extra substance (video, movement, 3D or other) arranged ahead of time to show up on the head of the marker. So, if the marker images is arranged effectively, marker-based AR content gives quality encounters and following is entirely steady, the AR content doesn't shake and easy to use but The augmented reality content may not make sense in a certain context[26].
2. Marker less AR: Marker less AR works by filtering the general condition and there is no trigger photograph important to recover the enlarged reality content. Marker less AR solutions started as hardware packages like Google Tango, but have evolved into software-based SDKs that produce the same effect without the need for specialized equipment[24]. When the substance is set in a room, it is more adaptable than marker-based other options but AR content may not make sense in certain context[26].
3. Projection AR: This is perhaps the least difficult kind of AR which is the projection of light on a surface. Projection-based AR is engaging and intelligent

where light is blown onto a surface and the cooperation is finished by contacting the extended surface with hand[28].
4. Superimposition Based AR

3.5 Virtual Reality

Virtual Reality (VR) is the utilization of computer innovation to make a simulated situation. In contrast to the conventional user interface, VR places the client inside an encounter. Rather than review a screen before them, clients are submerged and ready to connect with 3D universes. By recreating whatever the number of senses as could be expected under the circumstances, for example, vision, hearing, contact, even smell, the computer is changed into a guard to this artificial world. Quite far to move toward authentic VR experiences are the availability of substance and unassuming handling power. VR will cause us to feel like we are there intellectually and physically. We turn our head and the world turns with us so the fantasy made by whatever world we are in is rarely lost.

3.5.1 Basics of VR Works

The essential subject of VR is recreating the vision. Each headset means to consummate their way to deal with making a vivid 3D condition. Each VR headset sets up a screen before eyes accordingly, eliminating any collaboration with this present reality. Two self-adjust focal points are commonly positioned between the screen and the eyes that alter dependent on singular eye development and situating. The visuals on the screen are delivered either by utilizing a cell phone or HDMI link associated with a computer. To make a genuinely immersive computer-generated simulation there are sure essentials a frame rate of least 60fps, a similarly equipped invigorate rate and least 100-degree field of view however 180 degrees is ideal. The edge rate is the rate at which the GPU can deal with the pictures every second, the screen invigorates rate is the pace of the presentation to deliver pictures, and field of view is the degree to which the showcase can uphold eye and head development. Both of these doesn't function according to the principles the client can encounter idleness for example an excess of delay between their activities and the reaction from the screen[27].
We need the reaction to be under 20 milliseconds to deceive the cerebrum which is accomplished by joining all the above components in the correct extent. Among the significant headsets accessible today, Vive and Rift both have 110-degree FOVs, Google Cardboard has 90, the GearVR has 96 and the new Google Daydream presents to 120 degrees. Concerning outline rate, both HTC Vive and Oculus Rift accompany 90hz presentations, while the PlayStation VR offers a 60hz display[27].

3.5.2 Types of VR

1. Semi-immersive VR: 4-D films might be the conventional utilization of semi-immersive VR, yet now the innovation is fundamentally utilized for instruction. Pilot test programs including a moving cockpit and recreated condition on screens permit pilots to prepare without the dangers of flying a genuine aeroplane. As

Figure 3.6: Virtual Reality[18]

VR innovation improves, engineers can deliver comparative frameworks for other complicated task or professions[29].
2. Fully-immersive VR: This type of VR is commonly used for gaming and other entertainment purposes in VR arcades or even in our home[23]. It gives clients the most practical experience conceivable, complete with sight and sound. The VR headsets furnish high-goal content with a wide field of view. Regardless of whether you're flying or battling the miscreants, we will feel like you're truly there.
3. Non-immersive VR: Non-immersive reproductions are frequently overlooked as a real kind of VR, sincerely it is a fact that it's normal in our regular day to day existences. The normal computer game is in fact considered a non-immersive augmented simulation experience. Consider it, you're sitting in a physical space, collaborating with a virtual one[23].

3.6 Data-Oriented Technology Stack

The DOTS is the aggregate name for Unity's endeavour at reshaping its internal structure in a manner that is quicker, lighter, and, more significant, upgraded for the current monstrous multi-stringing world[20].
Three segments form the DOTS:
1.The Entity Component System (ECS)
2.The C# Job System
3.The Burst compiler

3.6.1 ECS

An (ECS) engineering isolates character (elements), information (segments), and conduct (frameworks). The architecture centres around the information. Some benefits like More noteworthy adaptability when characterizing objects by mixing

and coordinating reusable parts. Wipes out the issues of long legacy chains with complexly intertwined usefulness and Advances clean plan by means of decoupling, exemplification, modularization, reusability.By utilizing ECS design WE can make shorter and less convoluted code.It is not as solidly characterized as different examples, for example, MVC. Frameworks conduct is ECS model is change starting with one state then onto the next.

3.6.2 Burst Compiler

It is a specific code-generator that arranges a subset of C# (frequently called High-Performance C# or HPC#) into machine code that is, more often than not, littler and quicker than the one that is created by a proportionate C++ code[20].

Chapter 4

Proposed Methodology

In this research, we introduce a method for augmenting real world objects in a virtual environment with the help of deep neural networks and game engines where the proposed system will contain three parts that are, acquisition, processing and visualization. The proposed system will ensure augmentation of real-world objects in a virtual environment in real-time.
Model Fitting is an accuracy measurement depending on how accurate results can be provided by a machine learning model compared to the datasets provided. Model fitting can be divided into three categories. A model is said to be overfitted when the resultant data almost completely matches the data provided. Whereas, the model is said to be underfitted when the accuracy level of the resultant data does not match closely enough.

4.1 Estimating Pose of Personalized Body Models

The proposed model requires a humanoid as a guideline. Human pose estimation in a frame is a rough calculation of the humanoid structure which is most suitable for visual perception which is carried out in the following two phases. In the first phase, a comparison is made between the 3D structure of the rendered model and the humanoid model. While the second phase correlates the joint positions of the humanoid which were evaluated with the help of neural networks.

4.1.1 Human Body Model

In our proposed model we are particularly interested in human body models that are automatically generated from a humanoid model consisting of appropriate geometric primitives[12]. The process can be executed in two steps which are human body scanning followed by model rigging. The body scanning process requires capturing images of the human body from four different angles while the humanoid stands in T pose in front of an RGBD [39]. After automatically registering the scanned data for reconstructing the human body, rigging takes place[13]. The rigging basically generates a morphable human model to fit the scanned model and later the attributes such as skeletal bone location, skinning deformation information are transmitted on to the scanned model. It is noteworthy to mention that these processes are automated and require less than 2 minutes to complete the entire task. The body model generated consists of 94 bones in total. These bones include parts of the

face, fingers etc. from which the joints of Neck, Spine, Left Collar, Right Collar,Left Shoulder, Right Shoulder, Left elbow, Right Elbow, Left Hip, Right Hip, Left Knee, Right Knee, Left Arm, Right Arm are our top priority. The generated model's three dimensional position can be illustrated with the help of 3 attributes and the quaternion-based illustration of global location requires four more attributes. So, including the joint angles, a total of 36D parameter vectors is used to illustrate the virtually generated model. However, there are some limitations to a few specific poses even though the joint angles are logical. These poses result in impossible body configurations.

4.2 Objective Function

E(h,o) is an objective function that distinguishes between the generated 3D model and actual humanoid. The E(h,o) contains two main terms. They are ED and EJ.

$$\mathrm{h}^{*} = \arg_{\mathrm{h}} \min \mathrm{E}(\mathrm{h}, \mathrm{o}) \tag{4.1}$$

E(h,o) consists of two terms, ED and EJ. Specifically,

$$E(h, 0) = w_D E_D\left(h, d^{\circ}\right) + w_J E_J(h, 0) \tag{4.2}$$

Here, the first term has been used to calculate the difference between the depth map of the rendered model and the depth map of the actual humanoid model. On the other hand, the second term has been used to calculate the deracination of the joint positions between the two in terms of both 2D and 3D representation.

4.2.1 Depth Term ED

The color image and depth map is achieved from the rendered model where the depth map can be directly compared to the depth map of the actual humanoid model. Closeness between the two depth maps is considered as a positive sign for our hypothesis to be in the right track.

$$\mathrm{E_D}\left(\mathrm{h}, \mathrm{d}^{\circ}\right) = 1/\mathrm{N_p} \sum_{P \in B} \mathrm{C}\left(\left|\mathrm{d_P^h} - \mathrm{d_P^{\circ}}\right|, \mathrm{T}\right) \tag{4.3}$$

Here, in equation (3), Ed is used for finding the absolute value of the depth differences within the box containing the human model. To forestall deceptive focuses/anomalies from influencing it to an extreme and to robustify the mistake term a clamping function is utilized. A value of 0 in RGBD camera depth reading speaks to absence of estimation due to materials with reflective properties or infrared obstruction, that are omitted in equation (3). Setting Np equivalent to the quantity of delivered model focuses is a terrible decision, since this elevates speculations that venture to just a couple of pixels in the picture, rather than theories that clarify all the accessible perceptions. To dodge this, Np is set equivalent to the quantity of focuses inside B that are near the profundity profile of the past arrangement. Along these lines, all theories for a specific casing share a similar standardization.

4.2.2 The Joints Location term EJ

OpenPose gives assessments (je) of the 2D areas of human joints, from which 3d estimations can be formed through manipulation of depth information. Furthermore, with a speculation h of H, the 3d joint areas can be assessed and rendered in camera view to get jh. Subsequently, an error term EJ2D is characterized that is used to penalize errors between the theorized (jh) and assessed (je) 2D joint areas:

$$E_{J2D}\left(h, c^{\circ}\right)=\sum_{i=1}^{nj} w_i \left\|j_i^n-j_i^e\right\|_2 \tag{4.4}$$

.
Again, EJ3D is characterized as the term for penalizing errors between theorized (jh) and assessed (je) areas of 3d joints:

$$E_{J3D}(\mathrm{h}, 0)=1/\mathrm{W}\sum_{i=1}^{\pi} \mathrm{w}_i \left\|\mathrm{J}_i^n-\mathrm{J}_i^*\right\|_2 \tag{4.5}$$

. The OpenPose can pinpoint some joints with more precision while some are localized with comparatively less precision like, the hip joints are less precisely localized compared to the knees and head joints. A weighting scheme is used in both 3D (Eq 5) and 2D (Eq 4) terms of the objective function to counterbalance such behavior of the detector. This helps to prioritize more accurate joints. For ankles and wrists wi=1.8 and for hips wi=0.2 is used. In Eq. (4) and (5),

$$W=\sum_{i=1}^{N} w_i \tag{4.6}$$

where nj denotes the number of all contemplated joints. So, the position term of the accumulated joints Ej, can be defined as:

$$E_J(h, 0)=\alpha E_{J2D}\left(h, c^{\circ}\right)+(1-\alpha) E_{J3D}(h, 0) \tag{4.7}$$

Here, the constant $\alpha = 0.97$ that was set experimentally, balances the 2D and 3D error terms which gives EJ2D a dominant part to play. Moreover, α also counterbalances for measuring EJ2D and EJ3D in different arithmetic scales.

4.3 Gory Detail of Detail Optimization

The residual between the point cloud and the capsule pose is

$$\mathbf{R}(\mathbf{s})=\mathbf{R}_{\mathbf{c}}(\mathbf{C}(\mathbf{s})) \tag{4.8}$$

where R(s) is the residual as a function of skeleton's states, R(c) is the residual as a function of capsule poses and C(s) is capsule poses as a function of the skeleton pose s The Levenberg Marquardt update is computed as a solution to the following:

$$\left(\mathrm{J}^{\mathrm{T}} \mathrm{J}+\lambda_1 \operatorname{diag} \mathrm{J}^{\mathrm{T}} \mathrm{J}+\lambda_2 \mathrm{I}\right) \delta=\mathrm{J}^{\mathrm{T}} \mathbf{R}_{(\mathrm{S})} \tag{4.9}$$

Here, the constant $\alpha = 0.97$ that was set experimentally, balances the 2D and 3D error terms which gives EJ2D a dominant part to play. Moreover, α also counterbalances for measuring EJ2D and EJ3D in different arithmetic scales.
Where J is a Jacobian Of R(s), λ_1 is Levenberg-Marquardt dumping factor and is λ_2 Levenberg's original damping factor (both λ are configurable parameters; eventually we will optimize the values over a datasets. The reason we have both is that rather than being able to, so to say, converge in peace" , the algorithm is used to continuously track a moving target).
TO compute J^T J and J^T $R_(S))$ efficiently, and to simplify the inner loops, the calculation is split into an on-GPU per-point calculation involving only the points and capsules but not the skeletal kinematics, and a cheap on-CPU calculation involving only the skeletal joints and the capstules but not the points. This is done by applying the chain rule to (1) as following:

$$\mathrm{J^T} = J_s^t \left(J_c^T J_c\right) J_s \tag{4.10}$$

$$\mathbf{J^T R}_{(s)} = J_s^T \left(J_s^T J_c^T\right) \mathbf{R}_{(s)} \tag{4.11}$$

where Jc is a Jacobian Of the function Rc , and Js is a Jacobian of the function $C(s)$. In all Jacobian calculations involving any orientation quaternion Q under the derivative, a product $(1 + ai + bj + ck)Q$ where $a, b, c = 0$ is substituted as Q and the derivatives are computed; corresponding components from are used to update Q as $Q' = (1 + a^i + b^j + c^k)Q$ (the Q' is then normalized to keep it from exploding). This works as a "rotational" derivative, similar to torque in physics. $J_c^T R_(s))$ and $J_c^T J_c$ are computed in CapsuleFitter::FitError, and returned as an M sized array of a structure with 6 x 1 and 6 x 6 matrices, representing individual capsule's parts of $J_c^T R_(s))$ and that capsule's block on the diagonal of . $J_c^T J_c$. The on-GPU compute kernel only computes unique nonzero components of the $J_c^T J_c$ (it is symmetric) which are converted into $J_c^T J_c$ matrix after being retrieved from the GPU. The whole J_c over all points is never stored; note that for any matrix A

$$\left(\mathbf{A^T A}\right)_{\mathbf{i,j}} = \sum_{\mathbf{k}} \mathbf{A_{k,i} A_{k,j}} \tag{4.12}$$

thus $J_e^T J_e$ can be computed for every point individually (each point corresponds 3 elements of the sum above) and then summed, without storing and loading a 3N x 6M sized temporary. Same is done for $J_c^T R_(s))$. Its which likewise can be computed per point and summed. Note that $J_e^T J_e$ is a symmetric square inclining network comprising of M 6 x 6 squares on the diagonal, a square for each case. J_s Is figured in CapsuleToSkeletonJacobian () by considering all guardians joints of each case. As the Jacobians are very inadequate, and enhancement should be possible (also to the enhancement depicted before in the GPU part) by putting away square frameworks and playing out a meagre lattice increase including just nonzero blocks. During profiling, this enhancement was prior discovered to be pointless on as of now focused on equipment (work area) however may be required for versatile or other low controlled frameworks, or on the off chance that general exhibition improves to where it becomes noteworthy, or if the quantity of joints is expanded. That improvement, while simple to execute, was not done in light of a legitimate concern for keeping the code more direct at this purpose of advancement.

4.4 Block Diagram

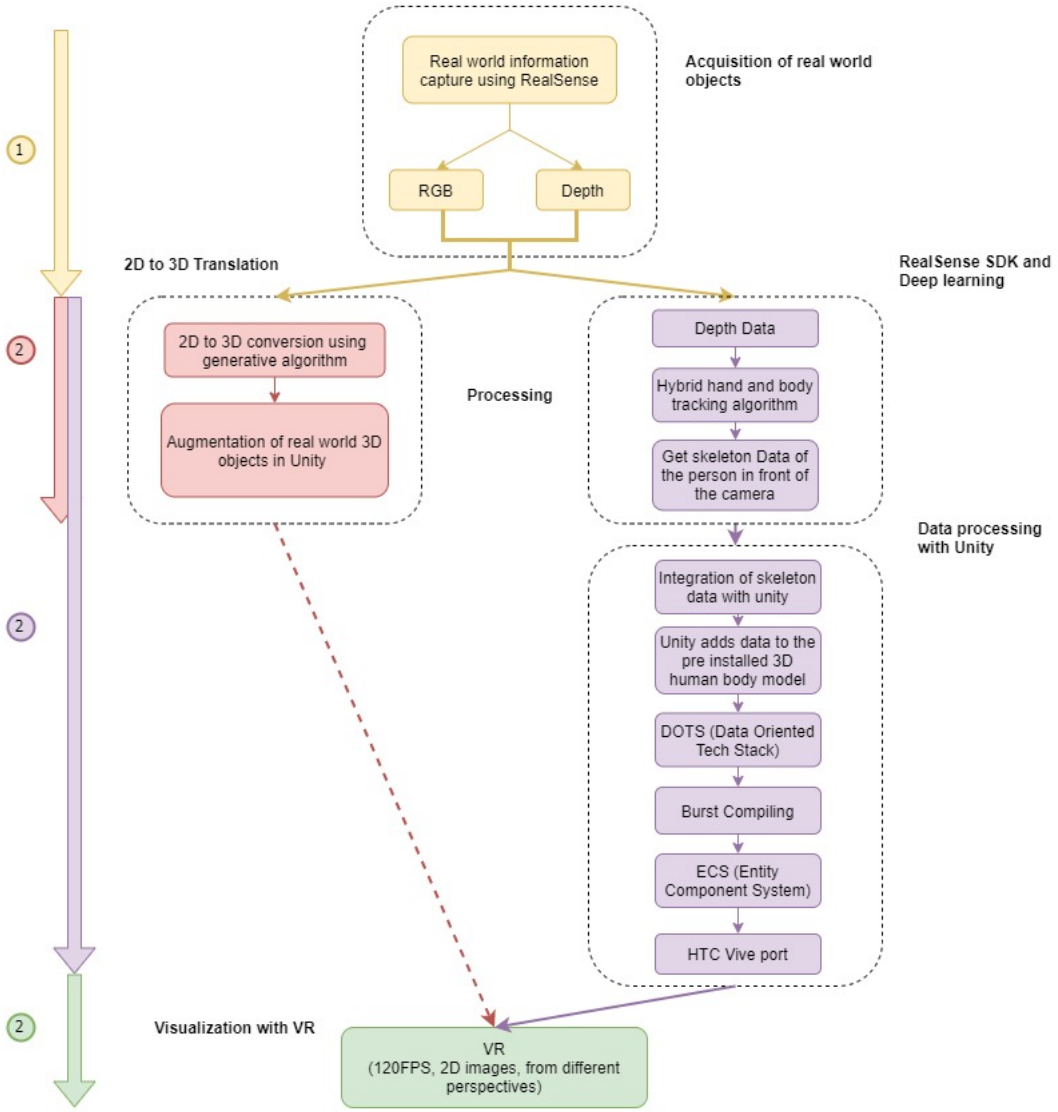


Figure 4.1: Workflow

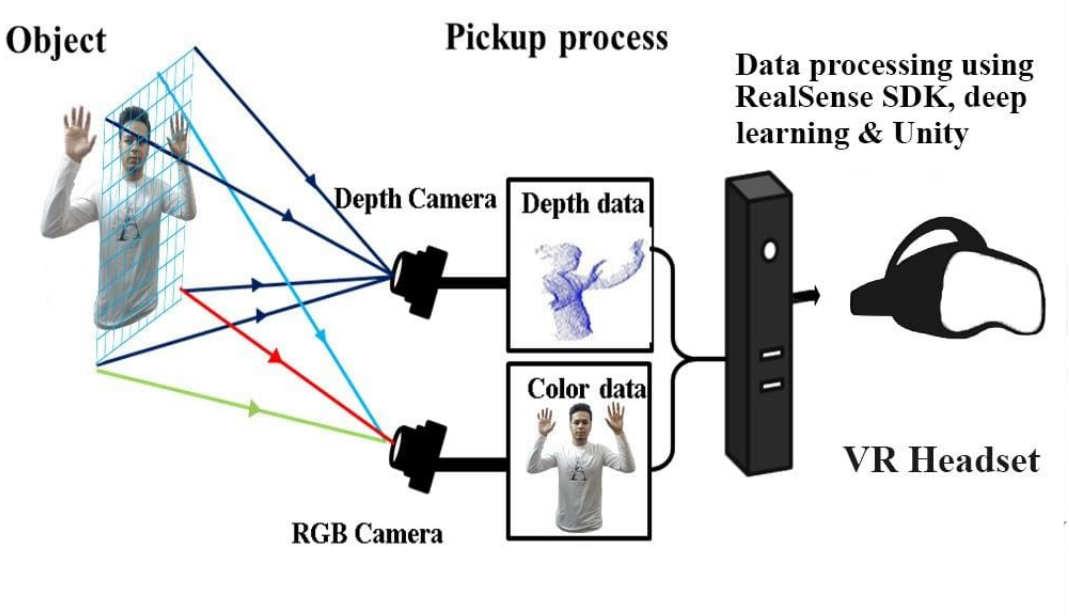


Figure 4.2: Data Acquisition

4.5 Components for augmentation in Unity Viveport

Expanding on the quick development and accomplishment of VR gaming, VIVEPORT highlights a wide scope of VR experiences. Viveport is the application store for VR contents that presents a wide scope of virtual reality experiences to enhance and motivate people in creating, connecting, shopping and discovering in virtual reality.

4.6 Unity 3d 2019.4(lts)

Unity is a cross-platform engine that supports building games for more than 25 platforms currently which includes smartphone, desktop, virtual reality, consoles etc. At present, almost more than 60 percent augmented reality and virtual reality contents are created using Unity. It is also worth mentioning that Unity platform easily connects to machine learning programs and the Unity Machine Learning Agents which is open-source software, uses trial and error that helps virtual characters to build creative strategies with reinforcement learning systems.
The virtual reality devices can be directly targeted by Unity's VR system with its base API and feature set that is compatible with handling multiple devices. The VE API is designed with minimalistic requirements that has the ability to grow along with the expansion of VR. After enabling VR in Unity some of the processes are automatically taken care of by Unity which include: Rendering all camera scenes directly to the head mounted display (HDM) and adjustment of view and projection matrices for field of view, head tracking and positional tracking so that the user's position and orientation matches closely to the orientation and position of the VR model. This process takes place before the frame is rendered providing good VR experience and also preventing users from experiencing nausea.

4.7 Steam Plugin

The SteamVR Unity plugin lets developers target a single API that can associate with all the renowned computer VR headsets. The advanced SteamVR Unity Plugin handles some of the major tasks for developers that include controller input management, VR controller 3d model loading and user hand estimation. Furthermore, the plugin includes an interaction system example to get the VR experience going.

4.8 GPU 1080Ti

The GTX 1080 Ti GP102 GPU contains 12B transistors that can operate at 1600MHz and can be further overclocked to 2000MHz. The GP102 has a built-in 6 graphics processing clusters consisting of 3584 CUDA cores, 88 ROPs and 224TMUs with a memory interface of 352-bit.

4.9 Data-Oriented Technology Stack

Using DOTS we are increasing the processing power. usually processing power assign in core only for one function but in our research we are assigning multiple cores only for one function. **Brust Compiling** is also a part of Dots. Brust compiling controls a lot of works at a single time and it compiles a lot of eventually. And the another one **Entity Component System(ECS)** where data, logic, method everything keeps in a different classes. The main motive is ECS compile every classes one by one

4.10 HTC Vive Pro

The HTC Vive Pro, designed for room-scale, immersive VR experience that was released in January, 2018 with upgraded features which include higher resolution displays of 1440x1600 for each eye, connected headphones, a secondary camera facing outward and a noise-cancellation analyzing microphone. All together, it is said to be a complete system with video, audio and accurate motion tracking. It's 615 ppi and 90 Hz refresh rate enables users to experience realistic movement, smooth gameplay and detailed graphics contents. The integrated microphones connected to the device not only provide active noise cancellation but also conversation modes and alerts. Whenever the user approaches the play area boundary, the device's built in Chaperone guidance system alerts the user and the physical elements are blended in by the front camera of the device. The Vive Pro can be used with both SteamVE 1.0 and SteamVR 2.0 Lighthouse base stations.

Chapter 5

Result and Analysis

5.1 Machine specifications

5.1.1 CPU specifications

The computer that has been used to perform the necessary steps of the research is a desktop computer. The processor this computer uses is Intel core i7-9700. It has total 8 cores, 8 threads, processor base frequency as 3.00 GHz (Max turbo frequency is 4.70 GHz), cache is 12 MB and bus speed is 8 GT/s. It has 16 GB of ram of memory type DDR4 and bus speed is 3200Mhz[36].

5.1.2 GPU specifications

The graphics processing unit (GPU) that has been used in the computer is Nvidia Ge-force GTX 1080 Ti. It runs on pascal GPU architecture and has 11 Gbps memory speed. It has total 3584 CUDA cores and has 484 GB per second memory bandwidth. This graphics card is fully compatible with virtual reality .

5.1.3 Images of setup

Here figure 5.1 shows the Intel RealSense camera we have used for data acquisition. Then figure 5.2 showcases the camera arrangement for the data acquisition part of our proposed model. These cameras are placed equally around the real-world subject.

Figure 5.1: The camera we have used

Figure 5.2: Camera arrangement

5.2 Result and Analysis

5.2.1 Result

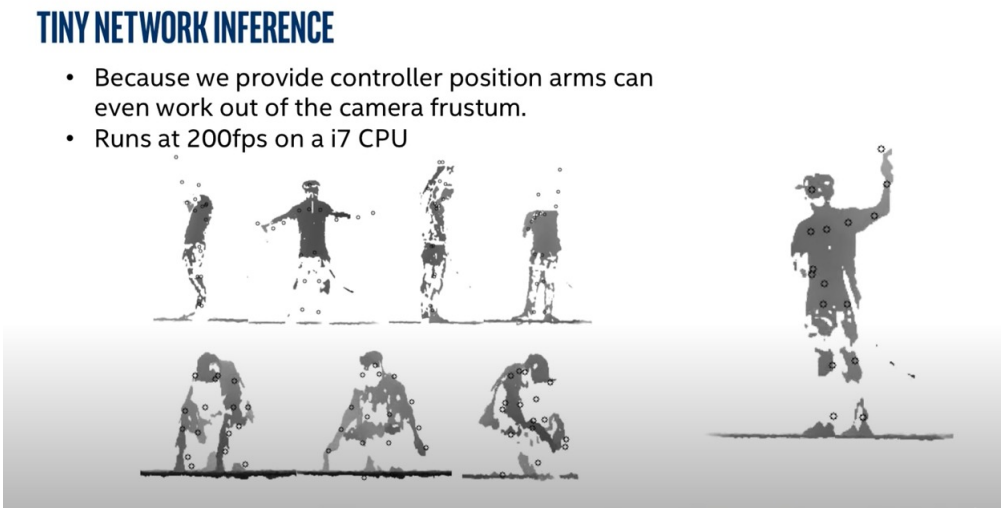


Figure 5.3: Depth Image

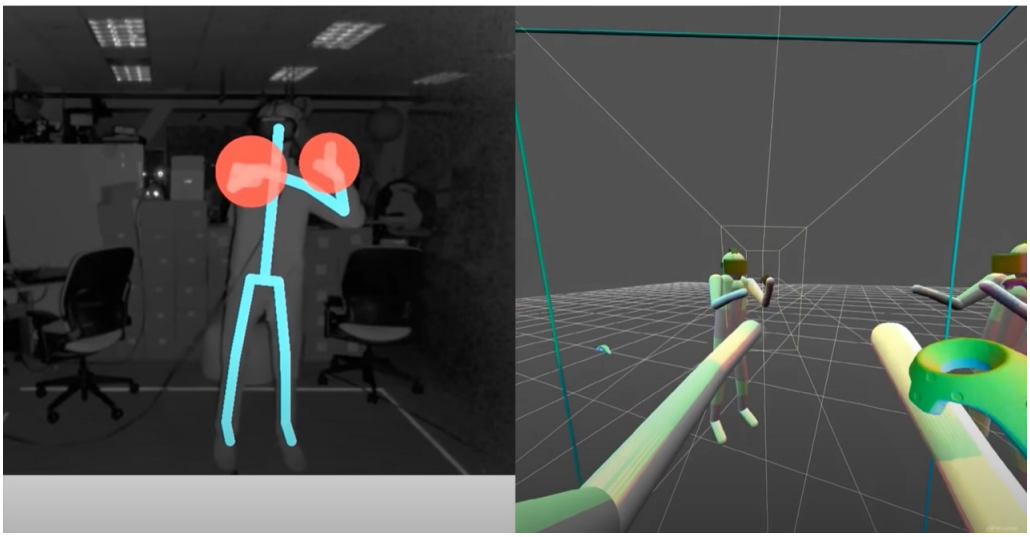


Figure 5.4: Skeleton data in unity

Figure 5.5: 3D model in unity

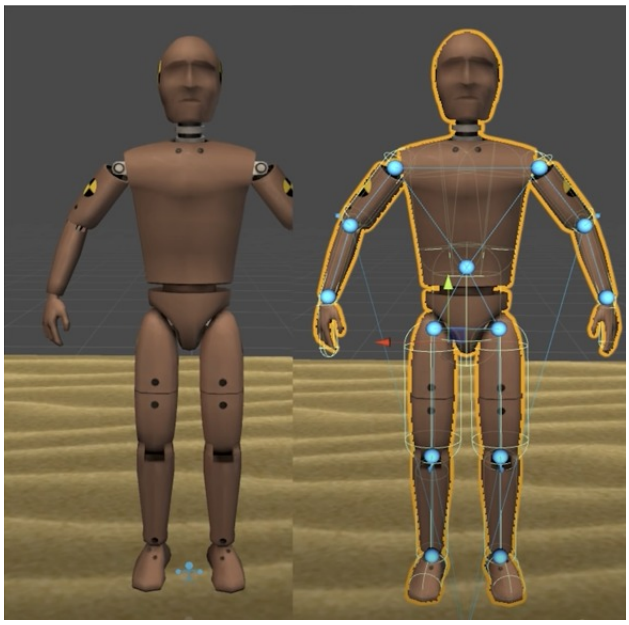


Figure 5.6: 3D modeling

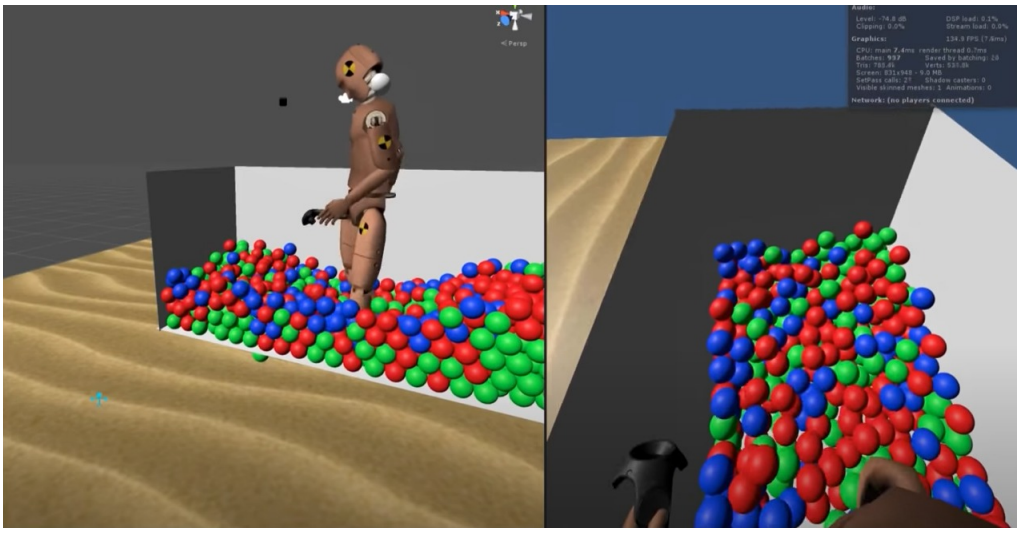


Figure 5.7: Final output in VR

Figure 5.8: Depth data to Skeleton

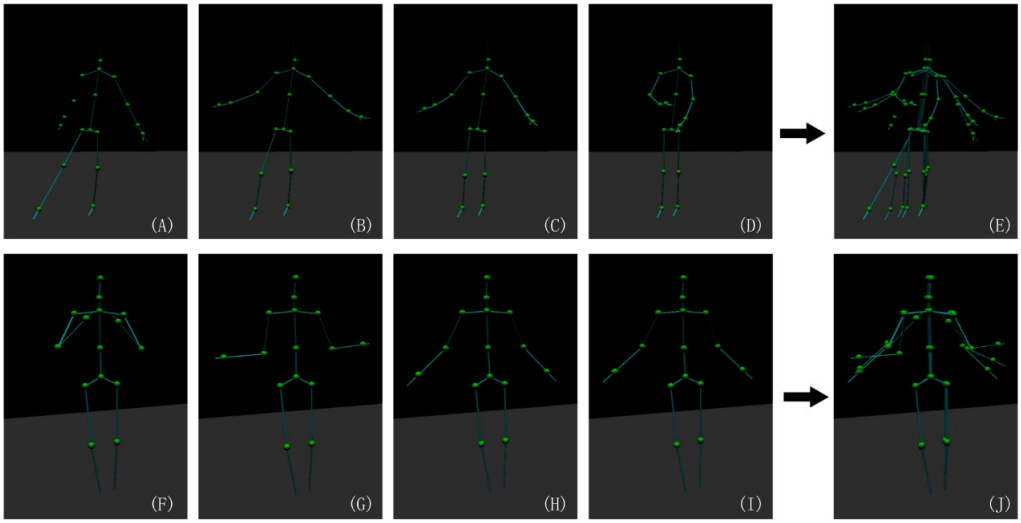


Figure 5.9: Data Picture

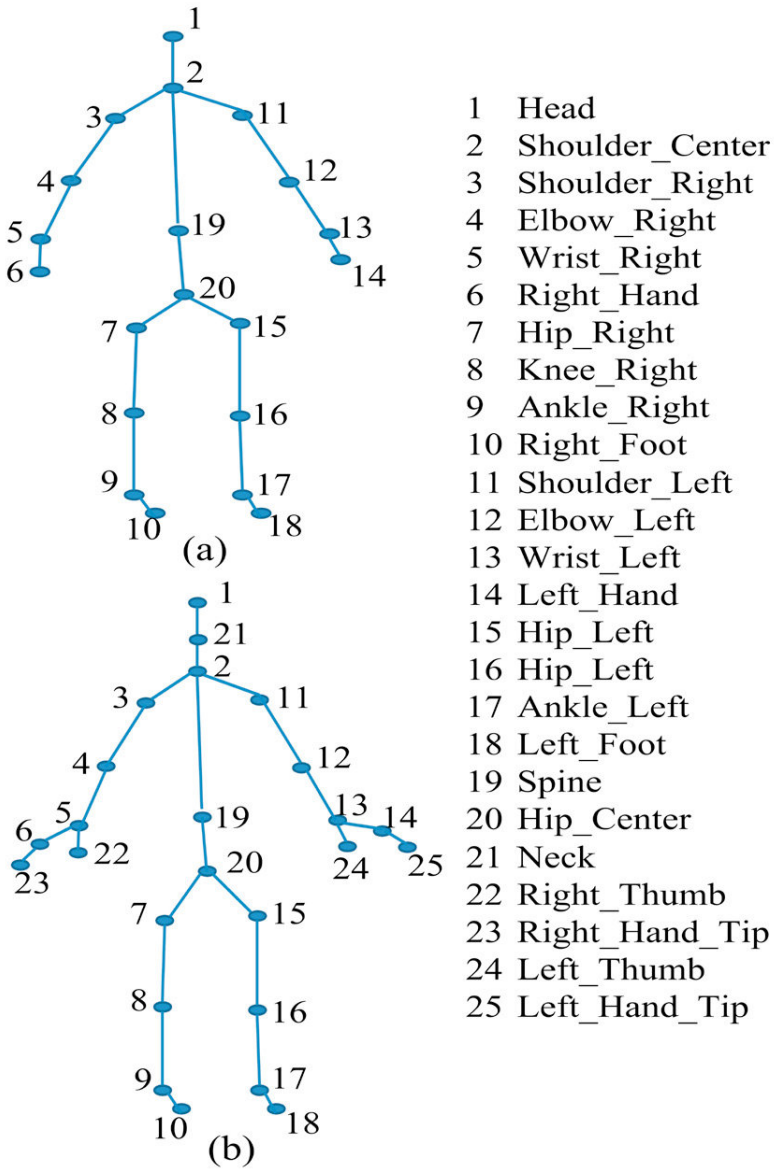


Figure 5.10: skeleton joint picture

5.2.2 Discussion

Here, figure 5.3 and 5.8 shows the depth image of the real-world subject after data acquisition. Figure 5.4 and figure 5.9 showcases the skeleton data that gets generated after using depth data in hybrid hand and body tracking algorithm. Figure 5.10 showcases the skeleton joint picture and how it is mapped. Figure 5.5 shows the predefined 3D model in Unity. Figure 5.6 shows the integration of the skeleton data with the predefined 3D model of Unity. The final output of the proposed methodology is showcased in figure 5.7 where the real-world subjects 3D model has been generated in a 3D environment which is mirroring gesture and posture of the subject in real-time.

5.3 Output devices

5.3.1 Depth Camera

The depth camera we are using here is Intel RealSense. This is the primary input device for acquiring data. Usually the images produced in digital cameras are in general 2D pixels. Those pixels have three type of values attached to it which are red, green and blue (RGB). The color varies from 0 to 255 for each of the RGB. For instance, white is (255,255,255) and blue is (0,0,255). These images are made by thousands of pixels. However, depth cameras have another value attached to the RGB value which is the depth value. The depth value means the distance of the camera from the object. These cameras provide each pixel with RGBD value where D stands for depth value[15]. The below picture shows two images where the actual 2D image has it's depth image right beside it. In the photographed image we see the 2D colors of each pixel which are RGB values. Then in the distance image we see that each color of the object represents a different value in distance of the object from the camera. Here, maroon is the closest and blue is the most distant from the camera.

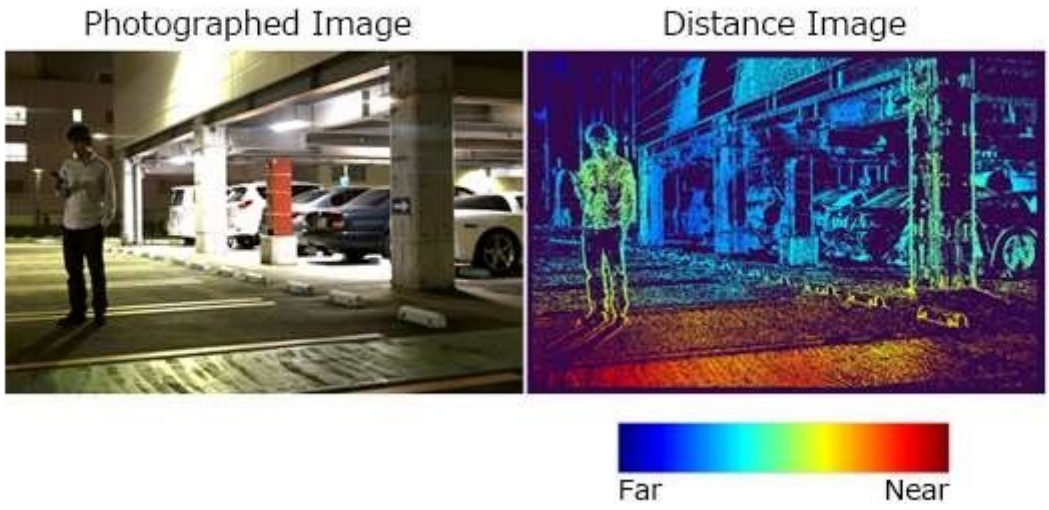


Figure 5.11: The camera we have used

5.3.2 VR

VR or virtual reality comes in the visualization part of our proposed model. This is the primary output device. By wearing the headset or VR devices a person will be able to interact in a simulated 3D environment. These devices can be of many kinds like screen attached goggles or tennis bat etc. For our research visualization, we are using HTC Vive pro. It will give a resolution of 2880x 1600 pixels[40]. Our goal is to generate 120 FPS visualization of the mirror 3D world through VR. It will only be 2D images from different perspectives but with 120FPS frame rate.

5.4 Development kits

5.4.1 RealSense SDK

Intel RealSense SDK is mainly a collection of pattern recognition and detection algorithm executions which are practiced on some specific interfaces. This library's goal is making these algorithms more accessible and moving developers' concentration towards algorithm details. So that they can put more contribution in innovation. In our research, when a person comes in front of the camera and gives input, the pixel depth data goes through hybrid hand and body tracking algorithm. This algorithm then tries to identify the skeleton by identifying for instance, ankle or knee etc. This whole process happens inside the RealSense SDK.

5.4.2 Unity

The development platform that has been used the most in our research is Unity. Most of the steps from the processing phase of our proposed model uses this development platform. Unity is a game engine and IDE. It can be used to make 2D, 3D, augmented reality and VR games or applications[31]. Unity is not only a game engine but also a strong IDE. We used Microsoft Visual Studio for coding for Unity.

Chapter 6

Conclusion

6.1 Conclusion

The implementation of the proposed model will unravel a range of various real-life solutions to the problems we never had any answer to. A real-time modeling of a real-life object in a 3D environment augmented inside VR can come handy in solving a lot of daily life issues. A lot of issues we are having in the present time as well can be treated better with the proposed model like meetings, virtual examinations, virtual classrooms, important seminars, games and many more. As the gaming industry is one of the most grossing industry in today's world, the current virtual gaming market is yet to be popular because of it's expensive tools that users need to buy. With the implementation of the proposed model a single depth camera will solve the issues and making VR games more interactive among gamers.

6.2 Future Work

Since this is the first stage of the model so for now, we have been able to make good progress in moving a predefined model in real-time with the person in unity. In our future development, we plan to not use any predefined model rather than very similar to distinct models' shape, size and volume centric models. For now the focus has only been to mirroring body parts of an avatar in real-time. In the future, we plan to make the avatar look like the person himself. Moreover, the branch of 2D to 3D translation using generative algorithm also falls under the future development of the proposed methodology. Which will most likely decrease the cost and increase it's performance and accuracy in getting modeling of a real-life object in a 3D environment augmented inside VR in real-time.

Bibliography

[1] J. K. Yan, N. S. Szabo, and L.-y. Chen, *Method and apparatus for texture generation - the singer company*, Sep. 1986. [Online]. Available: http://www.freepatentsonline.com/4615013.html.

[2] O. Faugeras and L. Robert, "What can two images tell us about a third one?", *International Journal of Computer Vision*, vol. 18, no. 1, pp. 5–19, 1996. DOI: 10.1007/bf00126137.

[3] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum, "Real-time texture synthesis by patch-based sampling", *ACM Transactions on Graphics*, vol. 20, no. 3, pp. 127–150, 2001. DOI: 10.1145/501786.501787.

[4] M. Zyda, *From visualsimulation to virtual reality to games*, Sep. 2005. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1510565.

[5] J. Toriwaki and H. Yoshida, *Fundamentals of three-dimensional digital image processing*. Springer Science & Business Media, 2009.

[6] Z. Khabazi, "Generative algorithms", *Accessed on*, vol. 3, 2011.

[7] L. A. Schwarz, A. Mkhitaryan, and D. Mateus, *Human skeleton tracking from depth data using geodesic distances and optical flow*, Dec. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S026288561100134X.

[8] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3d pose estimation from a single depth image", *2011 International Conference on Computer Vision*, pp. 731–738, 2011. DOI: 10.1109/iccv.2011.6126310.

[9] W. Fan and Z. Liang, *We apologize for the inconvenience...* 2012. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1237/2/022082/meta.

[10] T. Helten, M. Muller, and H.-P. Seidel, "Real-time body tracking with one depth camera and inertial sensors", *2013 IEEE International Conference on Computer Vision*, pp. 1105–1112, 2013. DOI: 10.1109/iccv.2013.141.

[11] C. Donalek, S. G. Djorgovski, and A. Cioc, "Immersive and collaborative data visualization using virtual reality platforms", *2014 IEEE International Conference on Big Data (Big Data)*, 2014. DOI: 10.1109/bigdata.2014.7004282.

[12] D. Michel, C. Panagiotakis, and A. A. Argyros, "Tracking the articulated motion of the human body with two rgbd cameras", *Machine Vision and Applications*, vol. 26, no. 1, pp. 41–54, 2014. DOI: 10.1007/s00138-014-0651-0.

[13] A. Feng, D. Casas, and A. Shapiro, "Avatar reshaping and automatic rigging using a deformable model", *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games - SA '15*, 2015. DOI: 10.1145/2822013.2822017.

[14] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks", *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, pp. 702–716, 2016. DOI: 10.1007/978-3-319-46487-9_43.

[15] J. F. Nunes, P. M. Moreira, and J. M. R. S. Tavares, "Human motion analysis and simulation tools", *Advances in Systems Analysis, Software Engineering, and High Performance Computing Handbook of Research on Computational Simulation and Modeling in Engineering*, pp. 359–388, 2016. DOI: 10.4018/978-1-4666-8823-0.ch012.

[16] J. Emspak, *What is augmented reality?*, Jun. 2018. [Online]. Available: https://www.livescience.com/34843-augmented-reality.html.

[17] natsu6767, *What are generative learning algorithms?*, Jul. 2018. [Online]. Available: https://mohitjain.me/2018/03/12/generative-learning-algorithms/.

[18] T. Olavsrud, *How virtual reality is helping embry-riddle transform air crash education*, Jun. 2018. [Online]. Available: https://www.cio.com/article/3280945/how-virtual-reality-is-helping-embry-riddle-transform-air-crash-education.html.

[19] N. Zhang and Y. Zhao, "3d recognition based on ordered images reconstruction", *MATEC Web of Conferences*, vol. 232, p. 02 045, Jan. 2018. DOI: 10.1051/matecconf/201823202045.

[20] S. L. - and V. D. -, *What is unity's new data-oriented technology stack (dots)*, Dec. 2019. [Online]. Available: https://hub.packtpub.com/what-is-unitys-new-data-oriented-technology-stack-dots/.

[21] Y. Chen, Q. Wang, H. Chen, X. Song, H. Tang, and M. Tian, "An overview of augmented reality technology", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1237, 2019, p. 022 082.

[22] A. C. N. Matcha, *An introduction to generative deep learning*, Oct. 2019. [Online]. Available: https://medium.com/analytics-vidhya/an-introduction-to-generative-deep-learning-792e93d1c6d4.

[23] B. Poetker, *What is virtual reality? (+3 types of vr experiences)*, Sep. 2019. [Online]. Available: https://learn.g2.com/virtual-reality.

[24] S. Schechter, *What is markerless augmented reality?*, Apr. 2019. [Online]. Available: https://www.marxentlabs.com/what-is-markerless-augmented-reality-dead-reckoning/.

[25] C. Software, *Supervised vs. unsupervised machine learning*, May 2019. [Online]. Available: https://medium.com/@chisoftware/supervised-vs-unsupervised-machine-learning-7f26118d5ee6.

[26] G. Zvejnieks, *Marker-based vs markerless augmented reality: Pros amp; cons: Overly app*, Dec. 2019. [Online]. Available: https://overlyapp.com/blog/marker-based-vs-markerless-augmented-reality-pros-cons-examples/.

[27] Anurag, *How vr works? know the technology behind virtual reality*, Aug. 2020. [Online]. Available: https://www.newgenapps.com/blog/how-vr-works-technology-behind-virtual-reality/.

[28] J. Gupta, *4 main types of augmented reality applications you can build*, Aug. 2020. [Online]. Available: https://www.quytech.com/blog/type-of-augmented-reality-app/.

[29] K. Matthews, *7 types of virtual reality that are changing the future*, Apr. 2020. [Online]. Available: https://www.smartdatacollective.com/7-types-of-virtual-reality-that-are-changing-future/.

[30] A. Rai, s. anushka sharma, and P. Srivastava, *Supervised and unsupervised learning*, Sep. 2020. [Online]. Available: https://www.geeksforgeeks.org/supervised-unsupervised-learning/.

[31] A. Sinki, *What is unity? everything you need to know*, Jun. 2020. [Online]. Available: https://www.androidauthority.com/what-is-unity-1131558/.

[32] L. Tagliaferri, *An introduction to machine learning*, Sep. 2020. [Online]. Available: https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning.

[33] Techslang, *How does augmented reality work?*, May 2020. [Online]. Available: https://www.techslang.com/how-does-augmented-reality-work/.

[34] A. Williams, *A beginners practical guide: Apply k-means clustering algorithm*, Apr. 2020. [Online]. Available: https://levelup.gitconnected.com/a-beginners-practical-guide-apply-k-means-clustering-algorithm-d57295699e61?gi=aaa5313266ff.

[35] E. Altberg, *Us8601386b2 - methods and systems to facilitate real time communications in virtual reality*. [Online]. Available: https://patents.google.com/patent/US8601386B2/en.

[36] Anonymous, *Intel® core™ i7-9700 processor (12m cache, up to 4.70 ghz) product specifications*. [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/191792/intel-core-i7-9700-processor-12m-cache-up-to-4-70-ghz.html.

[37] L. Herda, P. Fua, and R. Plankers, "Skeleton-based motion capture for robust reconstruction of human motion", *Proceedings Computer Animation 2000*, pp. 77–83, DOI: 10.1109/ca.2000.889046.

[38] S. V. LINDEN, *Us20080262910a1 - methods and systems to connect people via virtual reality for real time communications*. [Online]. Available: https://patents.google.com/patent/US20080262910A1/en.

[39] A. Shapiro, R. Wang, and A. Feng, *Computer animation and virtual worlds*, vol. 3-4.

[40] *Vive pro: The professional-grade vr headset*. [Online]. Available: https://www.vive.com/eu/product/vive-pro/.