UNIVERSIDADE FEDERAL FLUMINENSE

EDER DE OLIVEIRA

FPVRGame: DEEP LEARNING FOR HAND POSE RECOGNITION IN REAL-TIME USING LOW-END HMD

NITERÓI

2019

UNIVERSIDADE FEDERAL FLUMINENSE

EDER DE OLIVEIRA

FPVRGame: DEEP LEARNING FOR HAND POSE RECOGNITION IN REAL-TIME USING LOW-END HMD

Thesis presented to the Computing Graduate Program of the Universidade Federal Fluminense in partial fulfillment of the requirements for the degree of Master of Science. Topic Area: Visual Computing.

Advisor:
Prof. D.Sc. Esteban Walter Gonzalez Clua

NITERÓI

2019

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

O48f Oliveira, Eder de
FPVRGame: DEEP LEARNING FOR HAND POSE RECOGNITION IN REAL-TIME USING LOW-END HMD / Eder de Oliveira ; Esteban Walter Gonzalez Clua, orientador. Niterói, 2019.
72 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2019.

DOI: http://dx.doi.org/10.22409/PGC.2019.m.89111907134

1. Visão computacional. 2. Realidade virtual. 3. Interface de usuário (Sistema de computador). 4. Reconhecimento de padrão. 5. Produção intelectual. I. Clua, Esteban Walter Gonzalez, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

EDER DE OLIVEIRA

FPVRGame: DEEP LEARNING FOR HAND POSE RECOGNITION IN REAL-TIME USING LOW-END HMD

Thesis presented to the Computing Graduate Program of the Universidade Federal Fluminense in partial fulfillment of the requirements for the degree of Master of Science. Topic Area: Visual Computing.

Approved on july de 2019.

APPROVED BY

Prof. D.Sc. Esteban Walter Gonzalez Clua - Advisor, UFF

Prof.[a] D.Sc. Daniela Gorski Trevisan, UFF

Prof.[a] D.Sc. Luciana Cardoso de Castro Salgado, UFF

Prof.[a] D.Sc. Luciana Porcher Nedel, UFGRS

Niterói

2019

Thank you, sir, my God!

Acknowledgments

I want to thank my parents, João Batista de Oliveira and Sebastiana Laudelina de Oliveira, my brother Edir de Oliveira, my sister Ellen Cristina de Oliveira, my son Thiago Silva de Oliveira and my girlfriend Cristiane Eliza Mainardi, for believing in me, even when I did not believe. I love you all very much!

I want to thank my friend Bruno Augusto Dorta Marques for helping me so much and for inspiring me by his example.

I want to thank my adviser Esteban Walter Gonzalez Clua and the teachers Cristina Nader Vasconcelos, Daniela Gorski Trevisan, and Luciana Cardoso de Castro Salgado, for their knowledge and above all for their patience with me.

I want to thank all my friends who participated with me in this master's degree, and here I will not mention names because I would not forgive me if I forgot at least one of you.

Resumo

Os Head Mounted Displays (HMD's) tornaram-se dispositivos populares, aumentando drasticamente o uso da Realidade Virtual, Mista e Aumentada. Embora os recursos visuais dos sistemas sejam precisos e imersivos, as interfaces ainda são semelhantes às usadas em computação convencional, tais como joysticks e controladores, indo contra a expressão natural do corpo. Este trabalho apresenta uma abordagem para o uso de mãos nuas para controlar um sistema imersivo a partir de uma perspectiva egocêntrica que é construída por meio de uma metodologia de estudo de caso proposta. Utilizamos uma arquitetura CNN do DenseNet para realizar o reconhecimento em tempo real, com uma precisão média de 97,89%, tanto de ambientes internos quanto externos, não exigindo nenhum processo de segmentação de imagens. Nossa pesquisa também gerou um vocabulário, considerando as preferências dos usuários, buscando um conjunto de poses de mãos naturais e confortáveis e avaliando a satisfação e o desempenho dos usuários. Nós demonstramos nossos resultados usando HMD's comerciais de baixo custo e comparamos nossa solução com métodos de última geração.

Palavras-chave: Reconhecimento de Poses de Mão, Rede Neural Convolucional, Aprendizado Profundo, Realidade Virtual, Interfaces do Usuário.

Abstract

Head Mounted Displays (HMDs) became a popular device, drastically increasing the usage of Virtual, Mixed and Augmented Reality. While the systems' visual resources are accurate and immersive, precise interfaces require depth cameras or special joysticks, requiring either complex devices or not following the natural body expression. This work presents an approach for the usage of bare hands to control an immersive system from an egocentric perspective and built from a proposed case study methodology. We used a DenseNet CNN architecture to perform the recognition in real-time, with a mean accuracy of 97.89%, from both indoor and outdoor environments, not requiring any image segmentation process. Our research also generated a vocabulary, considering users' preferences, seeking a set of natural and comfortable hand poses and evaluated users' satisfaction and performance. We demonstrate our results using commercial low-end HMDs and compare our solution with state-of-the-art methods.

Keywords: Hand Poses Recognition, Convolutional Neural Network, Deep Learning, Virtual Reality, User Interfaces.

List of Figures

List of Tables

List of Abbreviations and Acronyms

ANN	:	Artificial Neural Network;
AV	:	Augmented Virtuality;
CNN	:	Convolutional Neural Network;
DenseNet	:	Densely Connected Convolutional Networks;
ResNet	:	Residual Network;
FPV	:	First-PersonVision;
GPS	:	Global Positioning System;
GPU	:	Graphics Processing Unit;
GoogLeNet	:	CNN Based on the Inception Architecture;
HDR	:	High Dynamic Range;
HMD	:	Head Mounted Display;
HCI	:	Human-Computer Interaction;
IMU	:	Inertial Measurement Unit;
MR	:	Mixed Reality;
MSE	:	Mean Squared Error;
NUIs	:	Natural User Interfaces;
R-CNN	:	Region-based Convolutional Network;
ReLu	:	Rectified Linear unit;
ResNet	:	Residual Network;
SGD	:	Stochastic Gradient Descent;
SH	:	Spherical Harmonics;
ToF	:	Time-of-Flight;
VR	:	Virtual Reality.

Contents

Chapter 1

Introduction

Head Mounted Displays (HMD's) [41] are becoming popular and accessible, leveraging Virtual, Mixed and Augmented Reality applications to a new level of consumption. A considerable amount of these market is strongly attached to low-end devices, based on smartphones as displays and computing hardware, enhancing the possibility of users interacting with virtual worlds anytime, anywhere, whether to watch a movie, work or play games [18].

High-end devices, such as Oculus Rift [28] or HTC Vive [44], provide sophisticated interfaces controllers and tracking systems, allowing powerful and complex interactions with the virtual environment [20]. Due to the lack of these components, mobile-based systems must be projected to be more straightforward and in many cases less immersive solutions.

1.1 Problem Statement

Visual immersion achieved by the HMDs can generate high interaction expectation among the users. It is common to observe, at interaction time, that the user makes undesired body and hand gestures, moved by a natural body instinct.

This body interaction cannot be implemented with regular HMD joysticks and controllers. Thus, research has been conducted to offer a more natural engagement to users. For instance, several body and hand gestures recognition solutions are being presented in the last years, some of them using very different approaches than traditional joysticks: heart rate monitors [40], Coulomb friction model [11], acoustic resonance analysis [47] and even clothing that restricts joint movements [1].

The usage of bare hands is what seems to be the most natural and immersive solution, and some researchers are working on this [34, 36]. In this sense, precise and comfortable solutions still require some dedicated hardware, such as depth cameras, structured light-based systems, and even Inertial Measurement Unit (IMU) based hardware [38].

Depth cameras (RGB-D) have the capacity of delivering depth information for each pixel, making possible the use of different techniques for geometry reconstruction and estimation of inverse kinematics bones positioning [34]. In indoor and controlled environments the depth cameras perform very well and are being vastly used. However, depth sensors can generate noisy depth maps, presenting some limitations: restricted field of view and range, near-infrared interference (such as light solar) and non-Lambertian reflections, and thus cannot acquire accurate measurements in outdoor environments [32, 33]. When the cameras not fixed, these issues become more critical, and we option let's not associate low-end HMD with IMU-based hardware.

When the cameras are not fixed, these issues become more critical, and we opted to let us not associate low-end HMD with IMU-based hardware. Therefore, how to present a solution capable of performing the bare hand recognition, in real-time and with high accuracy, in indoor and outdoor, using ordinary cameras as input devices through low-end platforms?

Besides, defining Natural User Interfaces (NUIs) is not an easy task, but often when we think about user interfaces that are natural and easy to use, we think of user interfaces where the interaction is direct and consistent with our natural behavior. Too often, people think that if they use, for example, gesture interaction, the user interface will be natural. Nevertheless, it is not always true. If we consider something like multi-touch gestures, some gestures come naturally and intuitively, such as swiping with one finger you scroll through pages or you move content from one side of the screen to the other. In this case, the gesture itself corresponds to the action you are performing. Some gestures, though, require more learning such as a four-finger swipe to the left or right. When you swipe to the left or right with four fingers, you will switch from one app to the next. The four-finger swipe is not intuitive, and it does not come naturally to us [25].

Bill Buxton [4] says that NUIs exploit skills that we have acquired through a lifetime of living in the world, which minimizes the cognitive load and therefore minimizes the distraction. He also states that NUIs should always be designed with the use of context in mind. No user interface can be natural in all use contexts and to all users. So, rather than try to design NUIs that are natural for all users, we should focus any NUI we design

on specific users and contexts.

Bowman et al. [3] questioned naturalism in 3D interface saying that high levels of naturalism can enhance performance and the overall user experience, but moderately natural 3D UIs can be unfamiliar and reduce performance. Traditional, less natural, interaction styles can provide good performance, but result in lower levels of presence, engagement, and fun. Dealing with this trade-off between naturalism versus performance is still a challenge and few efforts have been reported about how to explore the design space in order to find the appropriate and natural interaction for a specific context of use.

So, the question remaining is if NUIs are not just created by using modalities that can be naturally translated into interaction commands in the interface, how can we define what they are? This work addresses such issue by using a user-centered design approach to build a hand pose vocabulary for 3D user interaction in an egocentric vision scenario.

1.2 Research Objective

The main objective of this dissertation is to present a solution capable to perform the bare hand poses recognition, in real-time and with high accuracy, indoor and outdoor, using ordinary cameras as input devices through low-end platforms. Besides, the solution should be providing a user-centered design approach to build a hand pose vocabulary for 3D user interaction in an egocentric vision scenario, that is natural and comfortable.

1.3 Contributions

The hand pose recognition is a Machine Learning problem modeled as a pattern recognition task. Given that the state-of-the-art algorithms for pattern recognition in images are based on Convolutional Neural Networks (CNN) [35], we have chosen to use this approach in our work.

We adopted three of CNN's based architectures (GoogLeNet, Resnet, and DenseNet [42, 10, 12]) , with which we conducted several training sessions using two different datasets until we get to the model used in our solution.

We present an interaction solution based on bare hand interactions, which perform the real-time recognition with 98% accuracy and can be executed indoor and outdoor using ordinary cameras as input devices through low-end platforms, such as smartphones.

One of the data sets was created explicitly for this work, containing approximately 59,000 RGB images of the hand in a First-Person Vision (FPV) [15], see Figure 4.2.

We present a user-centered design approach to build a hand pose vocabulary for 3D user interaction in an egocentric vision scenario.

To simulate a first-person vision navigation system (FPV), we developed the FPVRGame, a virtual reality environment, see Chapter 4.

The main contributions can be summarized as:

1. Implementation of a CNN-based method for recognition of user's hand poses captured from a first-person vision navigation system (FPV) perspective in any environment (indoor and outdoor) and without any background or lighting constraints;
2. Creation of an open dataset with approximately 59,000 images of hand poses from a first-person vision navigation system (FPV) perspective;
3. Design process for generation and assessment of a hand pose vocabulary by using the Wizard of Oz method;
4. Empirical evaluation of user's experience and performance, using the proposed hand posture recognition in comparison to the main interfaces available for HMDs, in both low and high-end systems;

In addition, while we attempted to solve FPV systems, our solution can be trivially extended to any other interaction paradigm, depending only on providing a new image dataset.

1.4 Applicability and Evaluation

The FPVRGame design and evaluation process involved three empirical studies. In Study One we specified a preliminary vocabulary containing the hand poses considered more intuitive to represent the actions of a character from an FPV perspective. In Study Two we evaluated the hand poses existing in the preliminary vocabulary and built a new and more comfortable vocabulary, capturing the images required to create the dataset. Finally, in Study Three we validated our results using a low-end HMD and a simple VR environment.

1.5 Dissertation Structure

This dissertation is organized as follows: Chapter 2 describes the related works. Chapter 3 presents our CNN based solution for hand posture recognition. Chapter 4 presents the FPVRGame design and evaluation process, i.e., hand poses vocabulary construction process; the comparison between the accuracy achieved by our method with other interfaces. Chapter 5 discuss. Finally, Chapter 6 and 7 present the conclusions of our work and future works.

Chapter 2

Related Work

In this chapter, we present previous works that are related to our proposed solution. We discuss the types of technologies involved in recognizing hand poses and compare our work to others who make up state-of-the-art.

2.1 Overview of Hand Pose Recognition

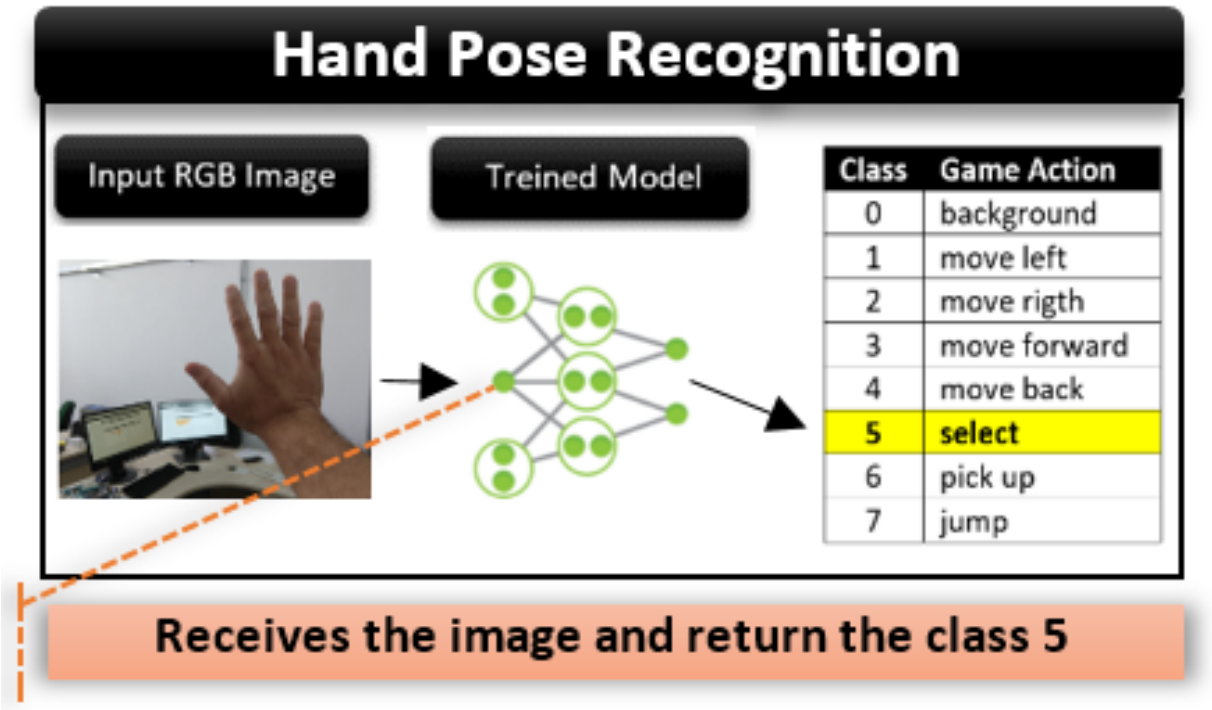


Figure 2.1: The process to convert a hand pose into a game command (action of selecting a coin).

The use of bare hands as a game controller requires to recognize the hand poses performed by the user during the game. Recognition refers to the whole process of identifying

the player's hands, making their representation possible, and converting a hand pose into a game command, such as the action of selecting an object, see Figure 2.1.

Two types of technologies promote this kind of interaction between humans and computers, vision-based and contact-based devices. Although there are several precise and functional interface devices for HMDs, we affirm that the use of bare hands is the most natural, intuitive and immersive [34].

Hand gestures can be classified using the following taxonomy: static and dynamic gestures. We chose to implement our solution through devices based on computer vision, using Convolutional Neural Networks (CNN's) [35], which in this work are responsible for classifying static gestures performed by the player and provide a numeric label, so that the actions of the game can be executed, see Figure 2.1 (Trained Model).

Because we choose to use low-cost HMDs, the recognition is made from RGB images that are captured by a simple camera coupled to a smartphone, and the images are obtained from a perspective of an egocentric view of the user, see Figure 2.1 (Input RGB image).

2.2 Datasets for CNNs

Training a CNN from scratch requires a significant amount of labeled data, being thus, we search in the literature for datasets containing images in First-Person Vision, also known as egocentric vision are available, such as [21, 43, 24, 48].

Li et al. [21] address the problem of recognizing the wearer's actions from videos captured by an egocentric camera. Its work encodes features hand pose, head motion, and gaze direction and using a similar pipeline with [45] tracks feature points in an input video with a time window of 6 frames, besides of extract a set of local descriptors.

Tewari et al. [43] introduced a dataset including the top view images of the palm and used a dedicated CNN architecture for hand pose recognition, its dataset of hand-pose was recorded using 3D Time-of-Flight (ToF) camera.

Molchanov et al. [24] used an algorithm for joint segmentation and classification of dynamic hand gestures from continuous depth, color, and stereo-IR data streams. For gesture recognition, it was used a network that employs a recurrent three dimensional (3D)-CNN.

Most of the datasets found contain dynamic gestures or use RGB-D images. Dy-

namic gestures recognition usually exploits spatiotemporal features extracted from video sequences. The evaluation of this type of method usually employs a sliding window of 16 frames or more [5], which introduces a significant input lag corresponding to the time between the 16 frames and the gesture recognition. Our work recognizes static gestures through a CNN, classifying hand poses from a single frame and allowing real-time recognition without the before-mentioned input lag penalty.

Since we did not find any dataset that represented all our gestures according to our context, we created a specific dataset to our context with a limited number of poses.

Tests performed in an egocentric dataset [48] show that our method has a competitive accuracy when taking account gestures similar to our hand poses vocabulary.

2.3 Similar Experiments of Hand Pose Recognition

Yousefi et al. [46] presented a gesture-based interaction system for immersive systems. Their solution makes use of the smartphone camera to recognize the gesture performed by the user's hands. The recognition process is based on matching a camera image with an image in a gesture dataset. The gesture dataset contains images of a user's hand performing one of the 4 available gestures. The images were recorded for both left and right hands under different rotations and a chroma key screen was employed to remove the background pixels. The construction of the dataset is labor-intensive, requiring the manual annotation of 19 joint points for each image in the dataset. At runtime, a preprocessing step is necessary to ensure that only the relevant data is fed to the gesture recognition system. This stage consists in segmenting the hand from the background and crop the image in the region of interest. The gesture recognition system performs a similarity analysis based on L1 and L2 norms to match the camera image with one of the dataset images. A selective search strategy based on the previous camera frame is used to reduce the search domain and efficiently recognize the gesture in real time.

The previous method requires extensive labor-intensive adjustments through the manual annotation of 19 joint points for each sample on the dataset creation process. Furthermore, the segmentation process requires manual adjustments based on the user's environment. As opposed to their approach, our method employs a dataset creation process that automatically annotates the images while the user is experimenting with the application. Furthermore, our recognition system works on raw input images and does not require background extraction, chroma key, and lighting adjustments.

Among several works, Son and Choi [39] proposed a hand pose detection approach that is capable for classifications based on raw RGB images. Their method recognizes three distinct hand poses employing a faster R-CNN, capable of identifying the region of interest and classifying one of the three possible poses. The dataset for training the network requires additional annotation for the palm position and fingertip. Their method aims to estimate the bounding box of the hands and identify which hand is in the camera field of view (left or right). In our paper, we consider that the game controller should work similarly for both hands, allowing left-handed and right-handed users to share the same experiences.

2.4 Hand-Based Interaction in Virtual Environments

We started looking at literature for hand-based interaction in virtual environments. We found many works presenting technological solutions for gestures interaction recognition [39, 21, 43, 46, 24, 48] and many others focusing on the usability and performance evaluation of the proposed hand gestures interactions [37, 22, 2, 31, 30]. We noticed that all of them focused on the assessment of the user experience and usability factors of using a specific hand-pose or gesture interaction. However, few of them described the design process used to choose those interactions, especially when those interactions cannot be directly mapped from natural users gestures.

For instance in the work of Rempel et al. [22] eighteen participants evaluated four different ray-casting hand gestures (index thrust, index click, palm thrust, and palm click) and three snapback thresholds while selecting 2D targets of different sizes. Dependent variables were mean time to select targets, a number of selections not completed, a number of incorrect targets selected, and subjective preference. The index thrust and index click gestures were preferred by subjects and had faster mean selection times and lower number of incorrect target selections.

In previous work, the purpose of Rempel and colleagues' study [30] was to develop a lexicon for 3-D hand gestures for common human-computer interaction (HCI) tasks by considering usability and effort ratings. Subjects (N = 30) with prior experience using 2-D gestures on touch screens performed 3-D gestures of their choice for 34 common HCI tasks and rated their gestures on preference, match, ease, and effort. Videos of the 1,300 generated gestures were analyzed for gesture popularity, order, and response times. The authors rated gesture hand postures on biomechanical risk and fatigue. A final task

gesture set was proposed based primarily on subjective ratings and hand posture risk. The different dimensions used for evaluating task gestures were not highly correlated and, therefore, measured different properties of the task-gesture match. A method is proposed for generating a user-developed 3-D gesture lexicon for common HCIs that involves subjective ratings and a posture risk rating for minimizing arm and hand fatigue. Results of this work were the start point of our first empirical study.

In another work [2], a study was conducted to explore the efficiency of hand tracking and virtual reality for 3D object manipulations in conceptual design. Based on existing research on conceptual design and hand gestures, an intuitive hand-based interaction model was proposed. An experiment on basic 3D manipulation shows that participants using a simple virtual reality and hand tracking interface prototype have similar performance to those using a traditional mouse and screen interface.

In Pirker et al. [31], they explore the Leap Motion controller as a gesture-controlled input device for computer games. Their work integrated gesture-based interactions into two different game setups to explore the suitability of this input device for interactive entertainment with focus on usability, user engagement, and personal motion control sensitivity, and compare it with traditional keyboard controls. The study proposes some gestures to interact with these games but did not discuss why they have chosen such gestures.

Chapter 3

Hand pose recognition solution

The hand pose recognition is the process of classifying poses of the user's hands in a given input image. We use the recognized pose to perform an action in an interactive game. Since we are using the player's hands like a game controller, the process must be robust to recognize the player's hands in multiple scenarios and different environments. It is important to have a consistent result in the hand recognition since a wrong classification would result in an involuntary movement in the game, potentially harming the user's experience.

The hand pose recognition is a Machine Learning problem modeled as a pattern recognition task. Given that the state-of-art the algorithms for pattern recognition in images are based on Convolutional Neural Networks [35], we choose to use this approach in our work.

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are learning algorithms that require a two-step process. A compute-intensive training step executed once, and a fast inference step, performed in the application runtime. Considering that our method aims to be executed in a mobile environment, the hand pose recognition must be executed in interactive time even on low spec mobile devices. This requirement makes the Deep Neural Network a suitable approach for our purposes.

The input of our method is a RGB image containing the user's hands (Table 4.1). The output is a probability distribution of the k possible classes. The classes are composed of the specified vocabulary described in Chapter 4, and an additional Background Class,

that represents the absence of the user's hands in the input image.

3.2 CNNs' Architectures

We adopted three CNN architectures for the hand pose recognition: GoogLeNet, Resnet, and DenseNet [42, 10, 12]. We discuss the different architectures in this subsection.

3.2.1 GoogLeNet

The GoogLeNet is a modular CNN based on the inception architecture [42]. It utilizes inception modules containing 1x1, 3x3, and 5x5 convolution layers plus an additional 3x3 polling layer. The modules are stacked 9 times by connecting the output of a previous module to the input of the next one.

The combination of multiple convolutional layers can lead to large computational complexity in the deeper layers of the neural network. Thus, a dimensionality reduction is performed by a 1x1 convolutional layer to reduce the number of filters for the expensive 3x3 and 5x5 convolution layers. Figure 3.1 shows the inception module with reduced dimensionality operation.

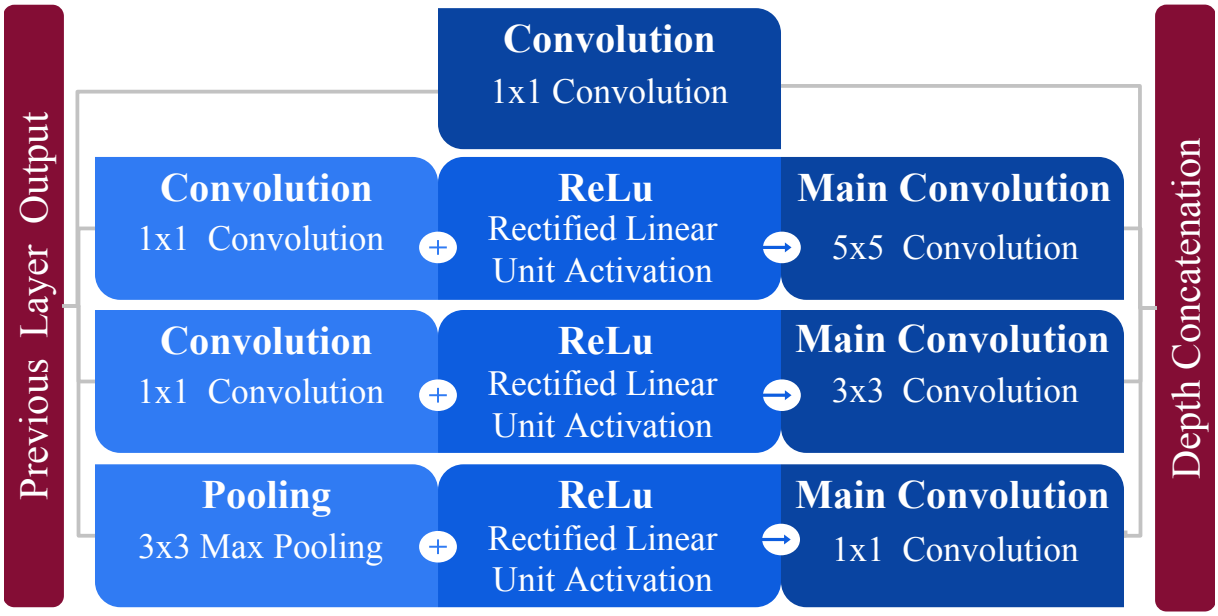


Figure 3.1: The inception module contains 1x1, 3x3 and a 5x5 convolution layers. The operations in the module are executed in parallel.

3.2.2 ResNet

The ResNet is a modular CNN based on the residual learning framework [10]. This CNN architecture makes use of shortcut connections to improve the classification accuracy while reducing the number of learned parameters when compared with GoogLeNet. The shortcut connection acts by merging the input with the output (Element-wise addition operation) of a ResNet module, resulting in a learnable identity mapping operation. Consequently, during training, the CNN is capable of flowing the gradients through the network skipping one or more modules.

The main layer in a ResNet module is a 3x3 convolution layer. To reduce the computational complexity of stacking multiple modules, a bottleneck design is improved. This design consists of inserting a 1x1 convolution layer, for dimensional reduction, before the 3x3 convolution, and a 1x1 convolution layer to restore the dimensions of the output in the building block. Thus, any module has the same feature size in the input and output but uses fewer parameters in the main convolution layer. An additional regularization is performed by the introduction of batch normalization layers in the ResNet [10]. Figure 3.2 illustrates the ResNet module with the bottleneck design, shortcut connection, and batch regularization.

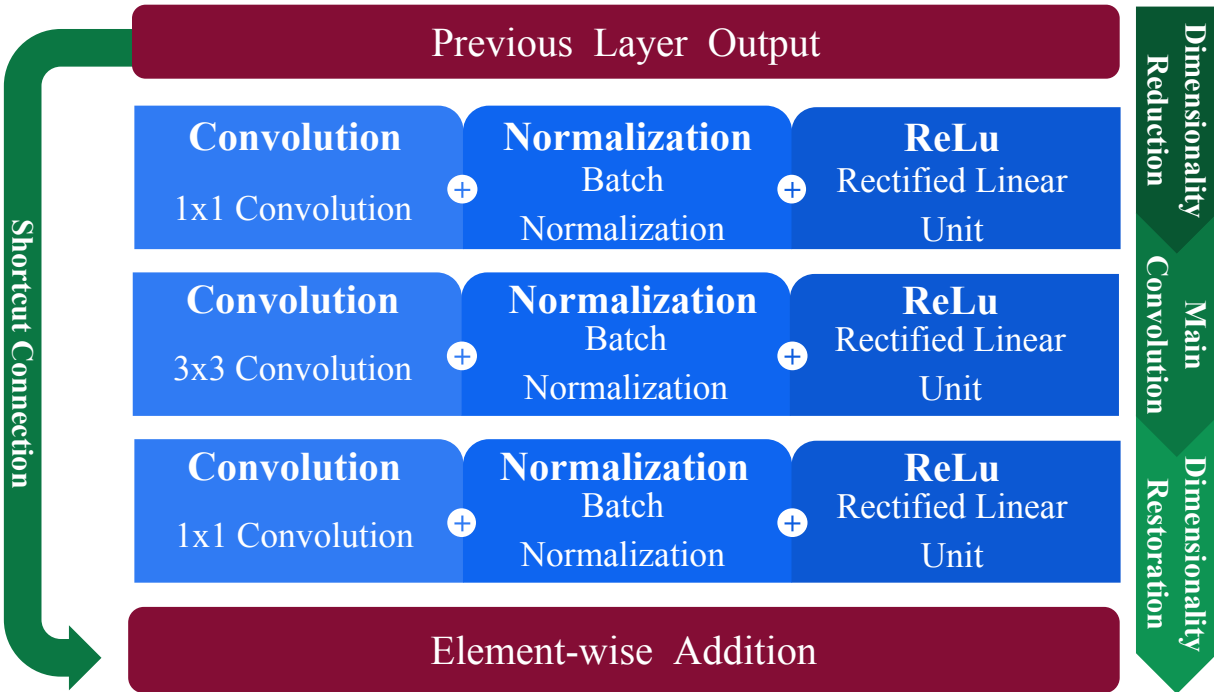


Figure 3.2: The ResNet module is a structure containing a 3x3 convolution layer in a bottleneck design with a residual shortcut connection.

Similarly to the GoogLeNet, the ResNet is constructed modularly by stacking mod-

ules, connecting the input of a module to the output of a previous one. In the cases where the feature map size are different between modules, a 1x1 convolution neural network is employed to adequate the dimensionality. In our experiments, we use a ResNet with 50 convolution layers.

3.2.3 DenseNet

The DenseNet is a CNN architecture that explores a simple connectivity pattern. Every layer is connected directly with each other. This pattern allows the network to achieve a similar accuracy with fewer parameters when compared to traditional CNNs [12]. This interconnection between layers also improves the flow of the gradients, thus helping the training of deeper network architectures. The connections also act as regulator for small training set sizes, effectively reducing overfitting during the training of the CNN.

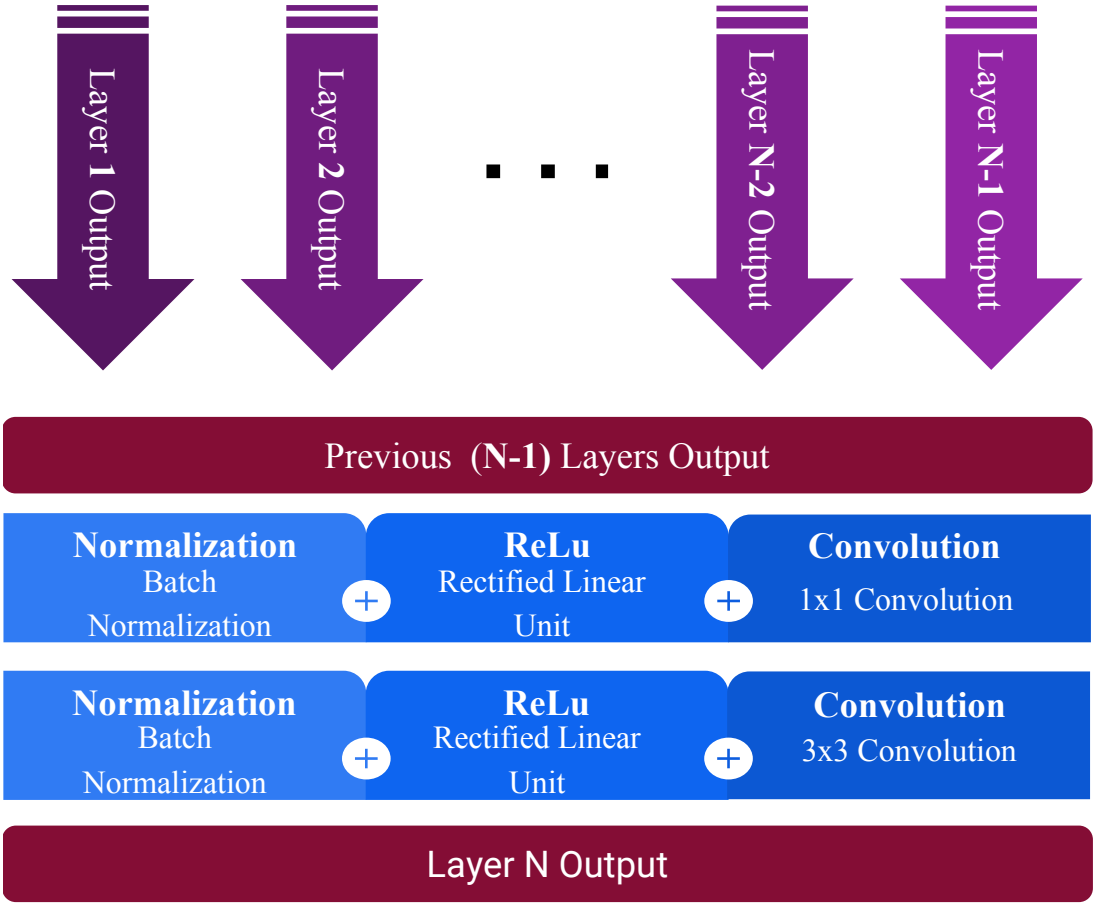


Figure 3.3: The DenseNet convolution block is a structure containing batch normalization and ReLu preactivated convolution layers. The input of the DenseNet block receives the output of all previous layers.

Every convolution layer in the DenseNet is preceded by a batch normalization and a ReLu activation function. The batch normalization layer act as a regulator that helps the training of the network, preventing overfitting of the training data. We arrange the convolution layer of the DenseNet in a structure named Convolution Block. This block contains a 1x1 convolution layer followed by a 3x3, as illustrated in Figure 3.3.

The construction of a DenseNet follows the modular approach where blocks are stacked to produce a deep CNN. A DenseNet module is a stack of densely connected convolutional blocks. The connections are made in a way that any layer of the network is connected to every other subsequent layer within the module. Figure 3.4 illustrates the interconnection of the Convolution Blocks within a DenseNet module.

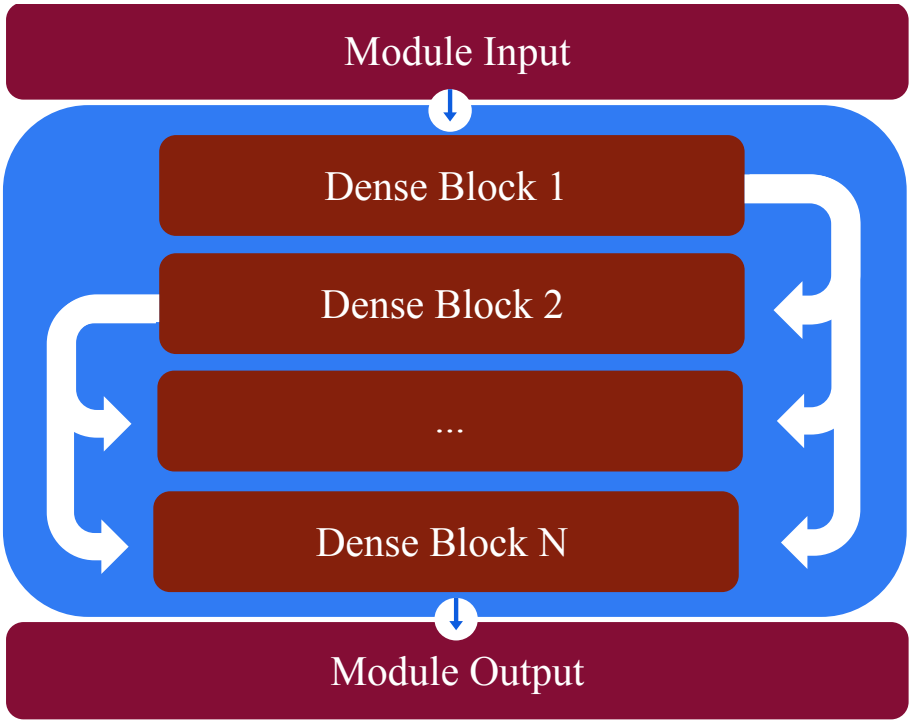


Figure 3.4: The DenseNet module stacks convolution blocks connecting the input of a block to all the previous blocks in the module. This interconnection improves the flow of the gradients between layers, allowing the training of deeper neural networks.

The DenseNet utilizes a hyperparameter k, referred to as the growth rate of the network. This hyperparameter defines the number of feature maps in the output of the convolutional layer of a DenseNet block. The DenseNet building blocks are connected by a transition layer composed of a 1x1 convolution and a 2x2 polling layer. In our experiments, we used a DenseNet with 121 layers and a growth rate of $k = 32$.

3.3 Datasets

We use two datasets in CNNs training. The dataset 1 is employed to precondition the CNN to recognize features related to human hands under different poses. The dataset 2 was explicitly created to meet the FPVRGame, see section 4.

3.3.1 Dataset 1 (DS1)

Figure 3.5: Example of DS1 images [19].

We use a pre-training dataset, containing 1,233,067 samples, taken from publicly available sign language datasets [19]. A sample in the dataset consists of a tuple (image, label) where the image portrays an interpreter performing a sign language gesture. Even though this dataset does not contain the correct label for our recognition system, the images of the dataset are employed to precondition the CNN to recognize features related to human hands under different poses.

3.3.2 Dataset 2 (DS2)

We created a second dataset specially tailored for our recognition system. The dataset consists of 58,868 samples captured in an indoor environment, comprising images of fifteen people (eleven men and four women). To improve our detection for both right-handed

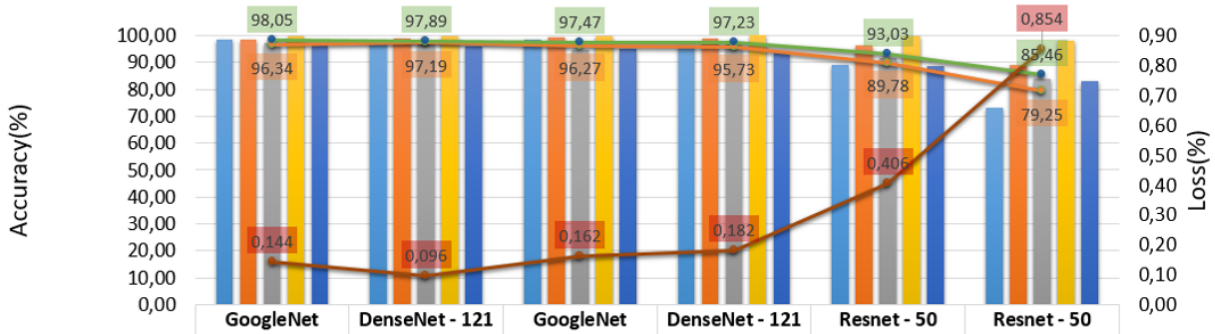
and left-handed users, we applied a mirror transformation in the images. The images were manually annotated with one of the seven classes that represent the possible actions of the game, or a class representing the background (Table 4.1).

A preprocessing step in both datasets ensures that the images have the same dimensions (256 x 256 pixels). A bicubic transformation was performed to resize images to adequate dimensions.

The process employed to acquire the images and annotation of the classes were performed during the Wizard of Oz study, described in Study Two in Section 4. The label annotation of the images was conducted during the acquisition of the images without the user's awareness. Resulting in obtaining natural images devoid of the conscious action of the user of effecting each hand pose, precisely the same as defined, as usually occurs when the collection is done with the conscious user.

3.4 CNNs' Training Results

Training a CNN from scratch requires a significant amount of labeled data; therefore, we trained three CNNs through a process known as fine-tune. We loaded a pre-trained model with weights adjusted to the ImageNet dataset[7]; then we fine-tuned our network to the DS2 dataset. Figure 3.6 shows a summary of the results obtained during the training.



	GoogleNet (DS1 and DS2)	DenseNet - 121 (DS2)	GoogleNet (DS2)	DenseNet - 121 (DS1 and DS2)	Resnet - 50 (DS1 and DS2)	Resnet - 50 (DS2)
Fold - 1	98,50	97,80	98,60	96,25	89,00	73,34
Fold - 2	98,28	98,93	99,33	99,04	96,38	89,02
Fold - 3	97,13	97,17	94,46	97,09	91,41	83,78
Fold - 4	99,90	99,86	99,82	99,99	99,92	98,07
Fold - 5	96,44	95,68	95,13	93,76	88,44	83,07
Mean validation	96,34	97,19	96,27	95,73	89,78	79,25
Mean test	98,05	97,89	97,47	97,23	93,03	85,46
Mean loss	0,144	0,096	0,162	0,182	0,406	0,854

Figure 3.6: Best results achieved during the training of CNNs, using the two data sets DS1 and DS2.

The ResNet architecture obtained a not satisfactory result, with a mean accuracy of 85.46% (test) and 79.25% (validation), with a mean error of 0.854%. Aiming to improve the accuracy of the classification, we performed a second experiment that exploits the features learned from the DS1 dataset.

The second experiment consists in, first, fine-tuning the CNN from the ImageNet dataset to the DS1 dataset, then fine-tuning the resulting model to the DS2 dataset. The result of this experiment for the ResNet architecture results in a mean accuracy of 93.03% (test) and 89,79% (validation) with a mean error of 0.406%. When compared to the first approach, we obtained a significant improvement in the mean accuracy of 8.85% (test) and 13.28% (validation) with a mean error of 0.406%.

While the ResNet highly benefit from the second approach, the GoogleNet and DenseNet obtained only a slight change in the mean accuracy when compared to the first experiment. The GoogLeNet test accuracy improved by a small margin (from 97.47% to 98.05%) while the DenseNet present a small decrease in test accuracy (from 97.89% to 97.23%).

Overall, the GoogLeNet obtained the highest mean accuracy of 98.05% on tests while the DenseNet, trained with the first approach, achieved the lowest mean error on the validation and a better mean accuracy distribution across the different classes. Furthermore, the training process was facilitated due to the usage of the first approach that does not require the finetuning to the DS1 dataset. The mean accuracy across multiple classes can be observed in the confusion matrix depicted in Figure 3.7. In our hand pose recognition system, we choose to use the DenseNet implementation due to less associated error across multiple classes and the relative uncomplicated single training process.

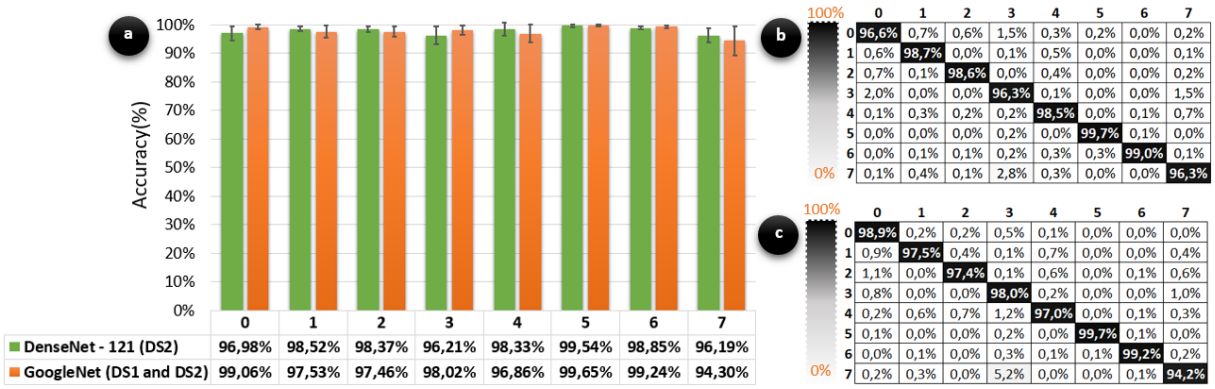


Figure 3.7: a) Mean class accuracy distribution (error bar: 95% confidence interval). b) DenseNet's confusion matrix. c) GoogleNet's confusion matrix.

3.4.1 Hardware Configuration and Training Parameters.

The CNN training was executed on a DGX-1 machine with the following specification: Intel Xeon E5-2698 v4 2.2Ghz, 512 GB DDR 4, 8 x NVIDIA P100 GPU. All the networks are trained using 4 GPUs adopting stochastic gradient descent (SGD) as our solver. The cross-validation is widely used to estimate the prediction error, thus we applied 5-fold cross-validation to our model, splitting the dataset into 5 distinct folds [8]. We achieved our best results running the tests for 30 epochs with batch size 96. The learning rate is set initially to 0.01 with the exponential decay (gamma=0.95).

3.5 Benchmarking

For validation purposes, we tested our trained CNN against a public available egocentric benchmark dataset, EgoGesture [48]. The dataset contains 2,081 RGB-D videos, 24,161 gesture samples totaling 2,953,224 frames. There are 83 classes of gestures, mainly focused on interaction with wearable devices. Because it is a data set different from ours, we have chosen a subset of gestures that are similar to the poses of our vocabulary. The mapping between the EgoGesture classes and our vocabulary classes is shown in the Table 3.1.

Table 3.1: Gesture mapping our vocabulary of hand poses to EgoGesture[48] gestures.

Our Vocabulary			EgoGesture		
	Class	Gesture		Class	Gesture
1	move left		66	thumb toward left	
2	move right		65	thumb toward right	
3	move forward		83	move fingers forward	
4	move back		67	thumbs backward	
5	select		29	number 5	

Our GloogleNet model, even though have never been trained with any of the images in the EgoCentric dataset achieved an accuracy of 64.3%. This result is superior to the

mean average accuracy of 62.5% in the VGG16 model presented by Cao et. al [5]. On one hand, we could improve our accuracy results by considering spatiotemporal strategies like appending an LSTM network to the output of our last fully connected layer, on the other hand, this introduction would increase the input lag in our application, thus making the model inadequate for VR applications.

Most of the errors associated with our model are the misclassification of gestures as background (class 0), as shown in the confusion matrix depicted in Figure 3.8a. This error is associated with the different nature of our training dataset and the tested dataset (video sequences in the EgoGesture vs single frames in our dataset). Frames at the two extremes (beginning and ending) of a video sequence in the EgoGesture dataset contains no identifiable gestures, for example, partially visible hands or the very beginning/ending of a gesture. To test this hypothesis, we tested our model by dropping the few beginning and ending frames of the video sequences. With 4 and 8 frames dropped, the model achieved a notable higher mean accuracy of 76.17% and 78.81%. The improvement in the model's accuracy and the confusion matrix (Figure 3.8b and c) confirm our hypothesis.

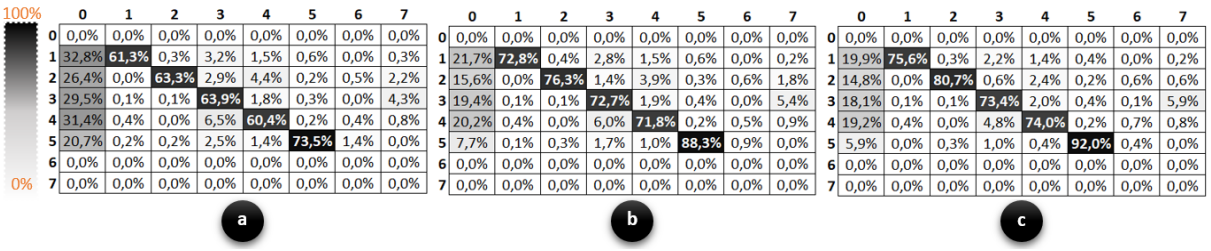


Figure 3.8: Confusion Matrix: a) 0 frames dropped, b) 4 frames dropped, c) 8 frames dropped. As for classes 6 and 7 we did not find similar gestures. Class 0 does not make its inference because it does not appear in the label of the base EgoCentric.

3.6 Inference Server Implementation

The recognition system is based on a client/server system. The FPVRGame, running on a smartphone Moto X4, act as a client that captures the HMD camera image and send them to the inference server application through a TCP/IP protocol. The server feeds the CNN with the received image and carries the recognized hand pose identification back to the FPVRGame. The inference server application (Figure 4.1c) was implemented with Python 3.6 using the Caffe framework [13] within an Intel® Core™ i7-7700HQ CPU @ 2.80GHz, 16GB RAM and NVIDIA GeForce® GTX 1050 Ti machine and running the DenseNet model performs the inference with an average of 28 milliseconds. Thus, the

inference process can run in real-time (35 fps) on any modern GPU enabled devices. Alternatively, it is possible to use our CNN model with third-party inference engines such as NVIDIA TensorRT [27], Clipper [6], and DeepDetect [14]. Moreover, the NVIDIA Inference Server [26] can be used for a cloud computing service solution.

Chapter 4

FPVRGame - Design and Evaluation Process

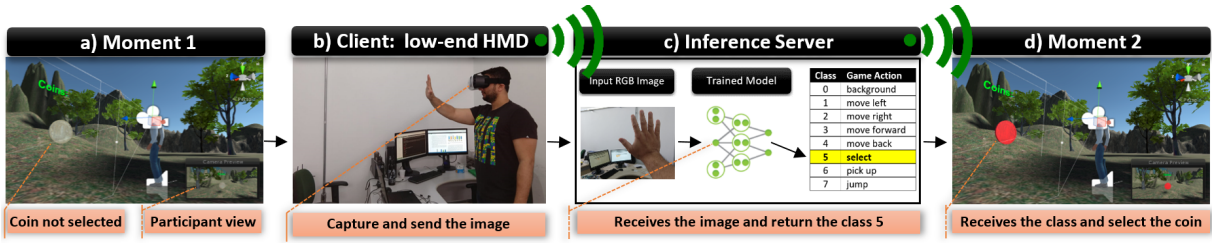


Figure 4.1: The FPVRGame is a VR environment developed to simulate an FPV game which allows the usage of the bare hand to control a character. This figure shows the action of selecting a coin and the process involved: the capture of the RGB image, the inference of CNN and the label send for execution of the action in the game.

FPVRGame is a VR environment developed to simulate a First Person View game, as illustrated in Figure 4.1. Its primary objective is to navigate through the scenario and collect as many coins as possible.

Providing input for the selection of objects and navigation in virtual, augmented, and mixed-mode reality can be done with hand-held controllers or hand gestures depending on the complexity and precision required. Freehand gestures have the advantages of eliminating the need for a game controller, not needing to see the controls of the controller.

Our game requires bare hands like a controller. For this, it captures the image of the player's hand and forwards the images to the Inference Server.

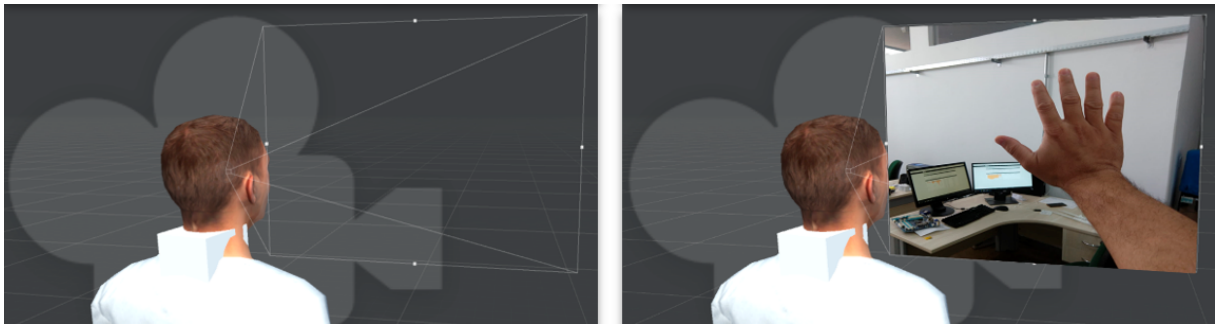
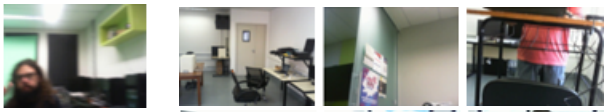
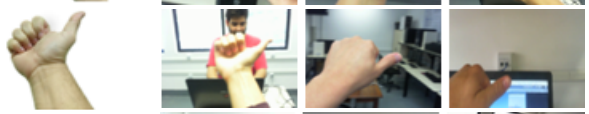
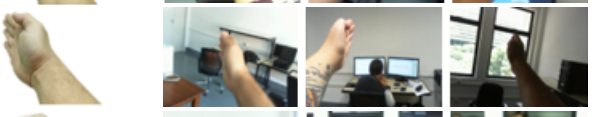
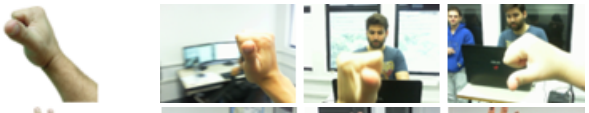
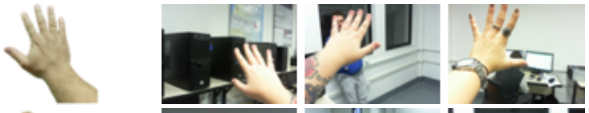


Figure 4.2: Camera angle and the field of view for hand positing.

As the camera is the single input device, the player has to position his hand inside the camera field of view (Figure 4.2). Due to this limitation, position the hand within the angle of the camera may require considerable physical effort and if the hand pose is not the most suitable, may cause pain.

Table 4.1 shows the vocabulary provided by FPVRGame, built based on two empirical studies, that offer a set of intuitive and comfortable hand poses, according to the experience of the users.

Table 4.1: Vocabulary of hand poses used for the player.

Class	Game action	Captured RGB images
0	background	
1	move left	
2	move right	
3	move forward	
4	move back	
5	select	
6	pick up	
7	jump	

Using the proposed vocabulary requires from the users to learn it as a dedicated movement because they need an understanding of the underlying system so as to understand the connection between the gesture and the action they are performing. In order to find which are the more appropriate hand poses for this context of use, we performed three empirical studies. The studies are described as follows.

4.1 Study One - Identifying Users' Preferences

Using an online questionnaire (APPENDIX A), we collected users' preferences about which would be the most appropriate hand poses to represent each player action.

4.1.1 Participants

We distributed the questionnaire to a group of users. A total of 173 people answered the questionnaire, 105 males, and 68 females, aged between 18 to 39 years (M = 25, SD = 4.06), right-handed 92.5% and 7.5% left-handed.

4.1.2 Procedures and Results

Participants were invited to imagine were using virtual reality glasses to simulate the use of the hand poses, to check how natural and comfortable they are. For each game's action, the questionnaire presented three images with different hand poses and a text field, so that the user could choose an option or describe a new hand pose, see APPENDIX A. Figure 4.3 shows the results obtained for the preliminary vocabulary of hand poses. The hand poses initially suggested in the questionnaire were selected from [30, 31].

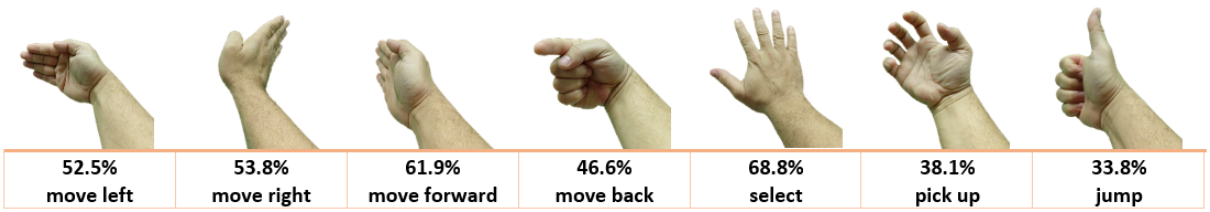


Figure 4.3: The preliminary vocabulary of hand poses (% of participants' choices).

4.2 Study Two - Formative Evaluation

This study aims to validate the preliminary vocabulary (Figure 4.3) investigating the users' experience in order to discard poses that are not intuitive and comfortable. We captured the participants' hand poses (with Wizard-of-Oz method) and generated the database for CNNs training, giving the participant the impression that they are using a ready-made game with the recognizing hand poses function [16].

4.2.1 Participants

We recruited 15 volunteers, 11 males and 4 females, aged 18 to 43 years (M = 26.46, SD = 6.94), all right-handed. To keep the magic of the Wizard of Oz, we chose the participants with little or no experience with games that use gesture recognition and participated in Study One.

4.2.2 Prototype

Figure 4.4: Prototype: a) participant performing the tasks, b) tutorial scenario, c) VR game with complete scenery and coins scattered. d) RGB images captured during the study to compose the DS2.

It was developed to simulate a VR game in the context of FPV where the player navigates the scenario and must collect as many coins as possible in the shortest time. It was composed of a tutorial (Figure 4.4b), formed by a small scenario with around the scene,

and a game (Figure 4.4c) with a bigger scenery and scattered coins. With the tutorial, the participant had the opportunity to become familiar with the game's actions and with the hand pose commands.

The prototype was designed to be used with the Wizard-of-Oz method with the following feedbacks: a frame that displays in real time the images captured by the camera of the mobile device and a set of buttons representing each one of the seven actions of the game. The frame should be used by the "wizard" to perform the game's actions synchronized with the hand's poses made by the participant. The set of buttons was used to give the impression that the recognition solution was working. Actually, they were triggered according to the decision of the "wizard", highlighting the action accordingly.

4.2.3 Setup

We used a controlled environment with: an HMD Oculus Rift DK1, a webcam (HD 720p LifeCam HD-3000 T3H-00011 MFT Microsoft) at the front of the HMD, both connected to a laptop running the prototype and with a turntable adapted to follow the player's movements. The "wizard" stands on the other side of the table, moving it as the player walks around the room and controls the game's character through the keyboard (Figure 4.4a). Participants needed to hand pose within the camera field of view, so their hands would be centered on the frame, and we could capture the participant's hands images, as shown in Figure 4.4d.

4.2.4 Task 1

The participant was invited to play the game using the set of hand poses that he/she has chosen in Study One. They started by following the tutorial, which advised them to perform all hand poses so that they could become familiar with the character's actions. The challenge proposed was simple: to collect five coins with no time limit. The purpose was to observe the participants' behavior and to stimulate their comments on how intuitive and comfortable it was.

After collecting all the coins, the participant was invited to play the complete game, which has the same features as the tutorial, but with a bigger scenario and with scattered coins. The goal was to collect as many coins as possible in 10 minutes.

We instructed the participant to think aloud while performing the assignments and recorded the entire study. After completing the task, we asked the participants the fol-

lowing questions:

Q1: Are the hand poses you performed natural and intuitive?

Q2: If, so, are they good representations of the game action?

Q3: What did you think of the hand poses you chose?

Q4: Did you feel any discomfort when using a particular pose? Which?

4.2.5 Task 2

The participant was invited to play again using the preliminary vocabulary of hand poses (Figure 4.3) defined in Study One. The same steps of Task 1 were repeated. After completing the task, we asked the participants to answer the following questions:

Q1: Did you enjoy playing using your hands as a control? From a grade of 1 to 5.

Q2: If you could change any pose, which changes would you make?

Q3: Would you like to repeat the game using the new hand poses?

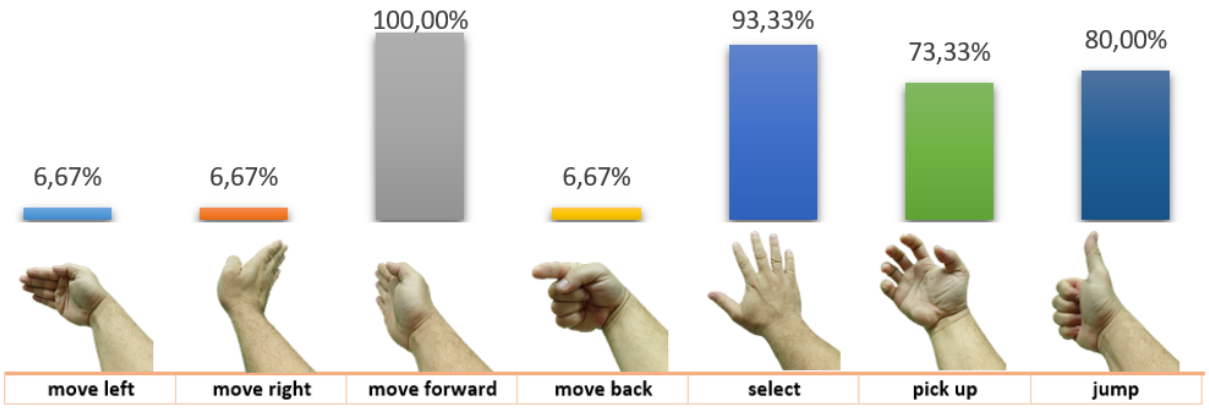


Figure 4.5: Final users' preferences (%) of the hand poses vocabulary after had performed the Task 2.

4.2.6 Task 3

We asked the participant to choose the hand postures that he considers appropriate to create a new vocabulary, being free to suggest other hand poses. Following, the participant was invited to play again with same steps of Task 1.

At the end of the tasks, we verified with the participant if the defined vocabulary has the most appropriate hand poses to this type of VR game. In case of positive answer, we proceeded to evaluate the recognition system. If not, the participant could repeat Task 3 until finding a definitive vocabulary that he considers appropriate.

Figure 4.6: Final users' preferences (%) of the hand poses vocabulary after had performed the Task 3.

4.2.7 Results and feedbacks

Figure 4.5 shows results from Task 2, where all participants evaluated the preliminary vocabulary. The hand pose "move left," "move right" and "move to back " had little acceptance. Only participant P11 reported that he could use them. All other participants, despite considering the intuitive poses, pointed out that felt muscle's pain and discomfort, even in the first minutes of the task.

Some selected comments about the preliminary vocabulary were:

"I felt pain when performing lateral movements, and difficulties remind me of the movements to jump and move forward";

"I want to change all the poses, except to move forward...";

"the movements to the left and the right, cause a certain discomfort";

"I found hand poses of the vocabulary very intuitive.".

Figure 4.6 shows that 4 of 7 hand poses of the preliminary vocabulary (Figure 4.3) have been replaced due to the discomfort reported by the participants. To replace actions, "move left" and "move right," 13 participants chose to use the thumb pointed left and right, respectively. To replace the action, "move back," 11 participants chose to use the thumb pointed back. All participants reported difficulties in positioning their hands

within the angle of the camera. Participants said that this disrupted the fun, because whenever they wanted to execute a command, there was a delay, due to the need to correct the hand's position.

4.3 Study Three - Summative Evaluation

A summative evaluation is focused on the outcome of a development process. This study aimed at assessing the participant's experience and performance when using our hand poses recognition solution. For that, we compare it with two other controllers (joystick and gaze) while manipulating a character in the FPVRGame. We decided to compare the hand pose interaction with these controllers once joystick is the most used commercial solution for games interaction while the gaze pointed based interaction has been recently used in VR mobile interaction [17] and [9]. Look gaze input means that you do not have to use a trigger. You just keep looking for a certain time on a target and then it gets selected.

Considering the feedbacks from Study Two, we raised the following hypotheses:

1. The perception of "easy to use" will be lower with the hand pose controller, when compared to gaze and joystick controllers.
2. The perception of comfort will be lower with the hand pose controller, when compared to gaze and joystick controllers.
3. The perception of affordance will be higher with the hand pose controller, when compared to the gaze and joystick controllers.
4. The perception of enjoyment will be higher with the hand pose controller when compared to gaze and joystick controllers.
5. The perception of the effectiveness of the hand pose interaction will be higher in indoor than outdoor environments.

4.3.1 Participants

Twenty subjects participated voluntarily in the study (15 males and 5 females). Their ages ranged from 21 to 37 years (M = 25.30, SD = 3.84). First thing to participate in the study, all subjects have to read and agreed with an Informed Consent Form. None of the

volunteers participated in the previous studies. Seventeen subjects reported being right-handed while the remaining were marked as left-handed. Ten subjects reported that they never used an HMD display or any tracking device before. However, all subjects reported having some experience with games. No correlation was found between the subject conditions and their performance during the experiment.

4.3.2 Setup

The equipment used in this study was a low-cost HMD consisting of VR BOX 2.0 glasses + Bluetooth remote controller (Joystick), a smartphone running the FPVRGame and a notebook running the Inference Server.

4.3.3 Procedures

Half of the participants played the game in an indoor environment while the other half played in an outdoor place. Participants played using all three game controllers alternating orders and were instructed to capture the coins as quickly as possible. After five minutes, we stopped the game and registered the number of coins collected.

The game actions using the gaze, are: look to an icon on the screen top for "move forward," look to coin per 1,5 seconds for "select" and look to coin per 1,5 seconds to "pick up" it. Using the joystick, the participant presses one button to "select" and other to "pick up." The hand poses used in this study were the same as those described in Table 4.1.

Each time the participant finished utilizing a specific game controller he was invited to fill out a questionnaire containing four questions:

Q1 - Easy to learn: I learned how to use this game controller easily.

Q2 - Comfort: I feel comfortable when using this game controller.

Q3 - Natural: I found natural using this game controller.

Q4 - Enjoyment: I enjoyed when using this game controller.

One more sentence was added to assess the feeling regarding the level of recognition of the hand poses interaction.

Q5 - Effectiveness: I found that recognizing poses was effective during my interaction.

We used a Likert Scale method for mapping the answers, with the following options: "strongly disagreed," "disagreed," "neutral," "agreed," or "strongly agreed." At the end of the experience, we asked the participant if they would like to change any of the hand poses used to play and to point out positive and/or negative aspects of his interaction. Finally, we asked the participants to rank the game controllers in order of their preference.

4.3.4 Results

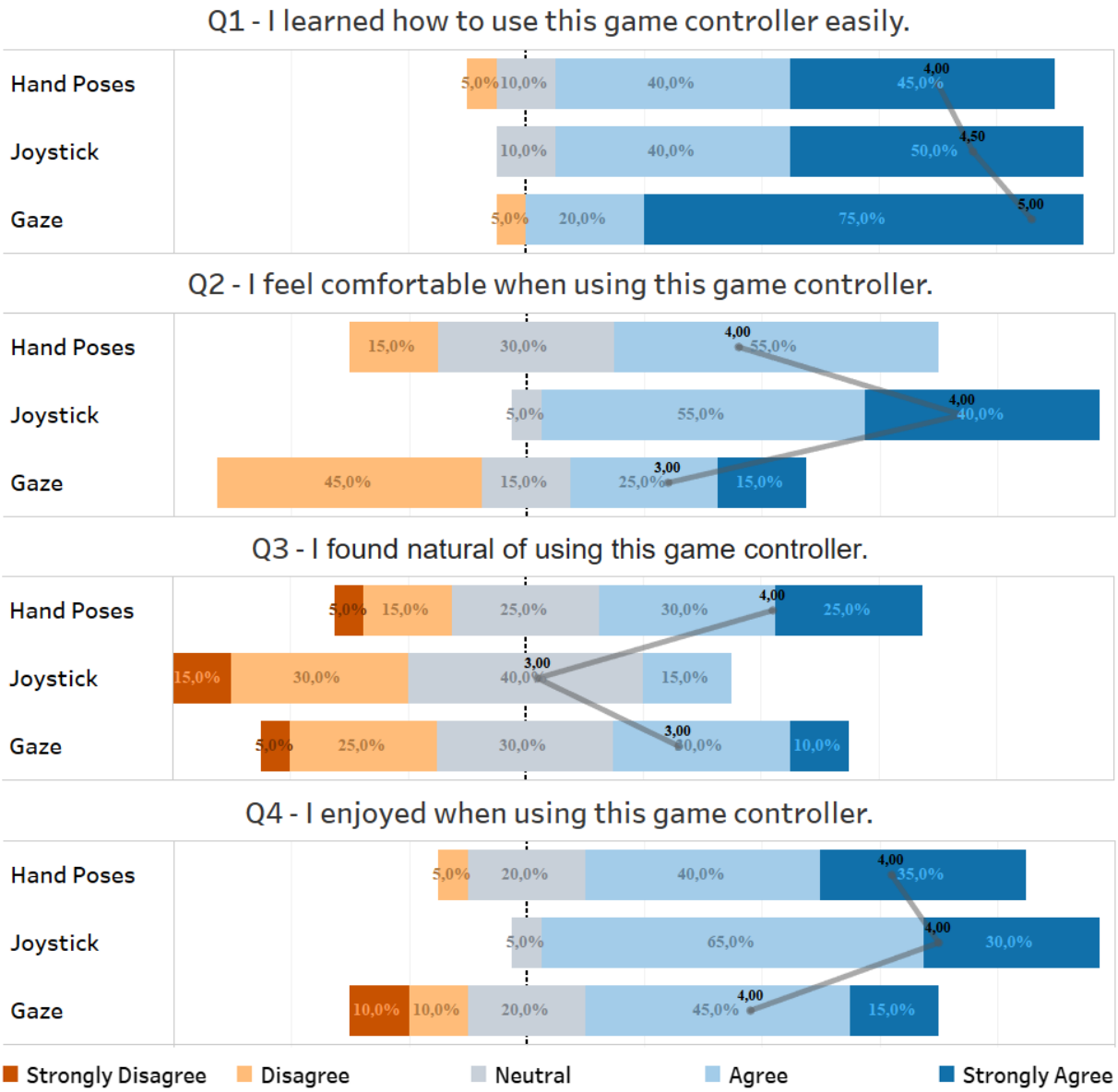


Figure 4.7: Perceptions of the participants considering all game controllers.

All statistical analyses were performed using IBM SPSS [1] with ($\alpha = 0.05$). The

[1]https://www.ibm.com/analytics/spss-statistics-software.

primary measures used in this study were the participant's feelings when using different game controllers (joystick, gaze and hand poses) to control a virtual character in the FPVRGame. Figure 4.7 shows a summary of the results for questions Q1, Q2, Q3, and Q4 for each game controller.

We find none significant difference between the game controllers (Friedman $X^2_{(2)} = 4.480$, $p = 0.106$) for the easy of learning question. Then our first hypothesis was not confirmed, and all controllers achieved a percentage of positive feelings equal to or greater than 85%.

For the item "Comfort," there was a significant difference between gaze and joystick controllers (Friedman $X^2_{(2)} = 16.033$, $p < 0.001$). The joystick achieved the highest percentage of positive feelings among the controllers (95%), being significantly higher than the gaze (40%). Then we did not confirm our second hypothesis where the hand pose interaction was expected to receive the worst comfort score.

For the item "Natural," there was a significant difference between the joystick and hand pose controllers (Friedman $X^2_{(2)} = 12.400$, $p = 0.002$). The hand pose achieved the highest percentage of positive feelings among the controllers (55%), being significantly higher than the joystick (15%). The hand pose interaction was expected to be more affordable than the joystick and gaze interactions.

For the "Enjoyment", there was no significant difference between the game controllers (Friedman $X^2_{(2)} = 8.041$, $p = 0.018$, with the multiple comparisons tests with p value adjusted (Gaze-Hand Pose p =0.207), (Gaze-Joystick $p = 0.144$), (Hand Pose-Joystick $p = 1$)). Then our fourth hypothesis was not confirmed and all controllers obtained a percentage of positive feelings equal to or greater than 60%.

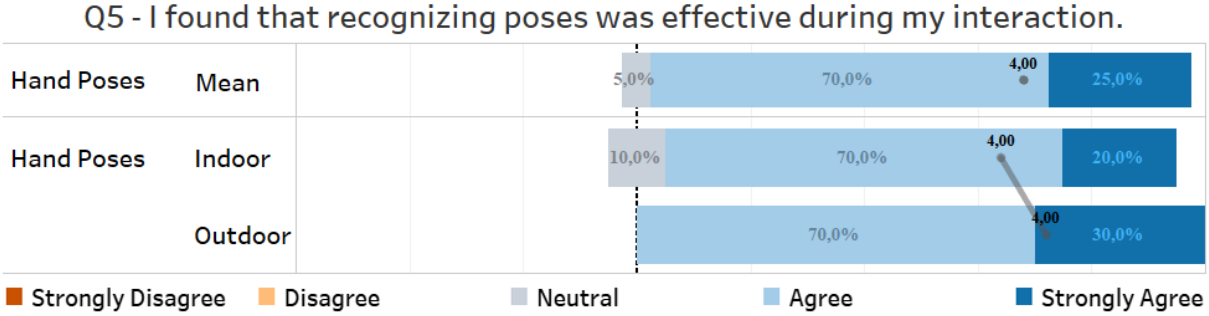


Figure 4.8: Participants' perception of effectiveness while using the hand pose controller, first row: all participants, second row: participants separated per group (indoor, outdoor).

For the evaluation of "Effectiveness" perception (Figure 4.8) the majority of the participants pointed out positive feelings regarding the effectiveness of the hand pose recognition

with only 5% of "Neutral" responses. In addition, we found no significant difference between the participants who performed the study in different environments (Mann-Whitney $U = 41.500$, $p = 0.423$) refuting our fifth hypothesis.

Concerning the performance of participants using the hand pose controller in indoor and outdoor environments we find none significant difference (t-test $t_{(18)} = 0.631$, $p = 0.536$). Coins collection in indoors environment had a result of 8.5 (SD=2.27) and in outdoors was 7.9 (SD=1.96). This result is in agreement with the evaluation of effectiveness perception.

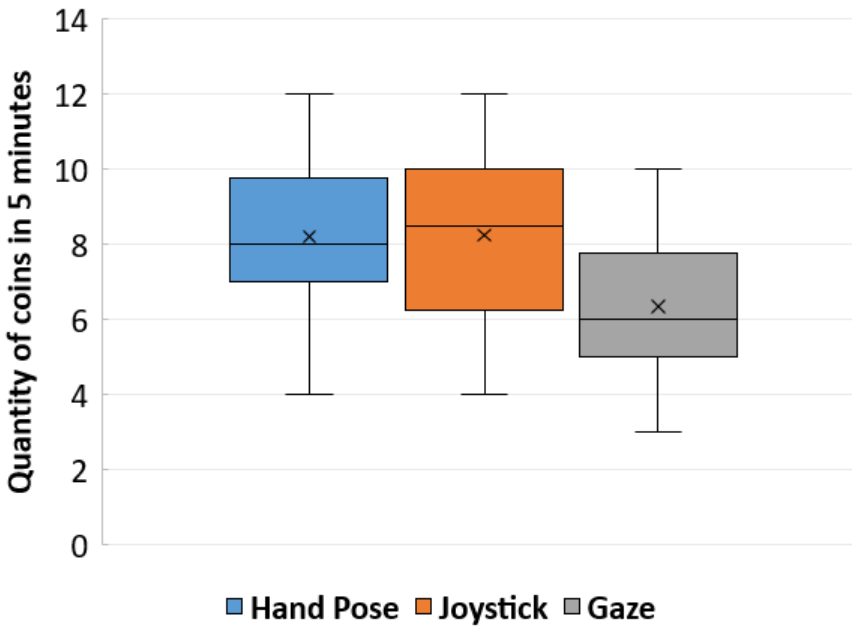


Figure 4.9: Quantity of coins per game controller: Hand Pose (M = 8.2, SD = 0.46), Joystick (M = 8.25, SD = 0.51) and Gaze (M = 6.35, SD = 0.43).

We also recorded the number of coins collected by participants using each game controller (see Figure 4.9). The Shapiro-Wilk test shows that the data follows a normal distribution (Hand Pose: $W = 0.970$, $p = 0.760$, Joystick: $W = 0.958$, $p = 0.499$ and Gaze: $W = 0.952$, $p = 0.396$). Thus, a One-way ANOVA with repeated measures ($\alpha = 0.05$) with *posthoc* and correction of Bonferroni was used ($F_{(2,38)} = 8.218$, $p = 0.001$. We note that the quantity of coins collected while using the hand pose controller was significantly higher when compared to the gaze controller. However, it was not different from the performance achieved while using the joystick.

Only one of the twenty participants reported having an interest in changing their hands during the interaction. The participants' preference about controllers was: hand pose with 9 votes, joystick with 8 votes and gaze with 3 votes.

We observed that the participants' preference of controllers followed the same behavior from participants' performance of controllers.

Chapter 5

Discussion

The Wizard-of-Oz technique used in Study Two showed to be adequate for validating the preliminary vocabulary achieved by Study One and contributing for a good final user's experience with the FPVRGame as discussed in Study Three.

We noticed that only 1 of 20 participants reported having an interest in changing the hand poses suggested. Concerning the ease of learning aspect, the hand pose interaction achieved positive feelings, similar to the other game controllers (see Figure 4.7). While most participants reported positive feedbacks saying: *"easy to use," "easy to interact," "easy to adapt"*, only 2 of 20 reported negative feedbacks: *"learning is slower," "I can not remember the commands of the hand"*.

In addition, the simulation with the wizard allowed appropriate conditions for recording the images and generating the data set. Before using the simulation as described in Study Two, we tried to ask for volunteers to perform some hand poses to be captured and used in the CNN's training. However, in practice, the results were not good and we assume that it was due to the robotic and not natural movements made by the voluntaries without causing oscillations in the poses. Even with a dataset composed only of images collected indoor, the CNN model was able to generalize the recognition of hand poses, and it works appropriately for both indoors and outdoors environments. Besides that, we observed that the performance of the hand pose interaction with the FPVRGame presented satisfactory result, similar to the joystick in the number of collected coins (Figure 4.9).

One limitation of FPVRGame is to use a single input device. The interaction occurs only when the user places his hand within the angle of the camera (Figure 4.7), which can cause some physical discomfort. Therefore, we evaluated the participants' comfort when

using hand pose and surprisingly such aspect was not a problem for the participants to enjoy and to achieve good performance.

We identified a significant correlation between the "Comfort" and "Enjoyment" perceptions (Spearman's $\rho = 0.654$, $p = 0.002$). It demonstrated that the higher the comfort is, the higher will be the pleasure. However, based on the percentage of positive feelings and the feedbacks from Study Three, where comfort was not the best neither the worst, we can assume that the level of discomfort does not prevent the user from feeling pleasure when using hand pose. Some of these positive comments were: *"very intuitive and fun," "It is more fun to use the body and does not need any external control"*. However, some negative observations were pointed, such as: *"I tried not to keep my arm straight," "I tire more"*.

Chapter 6

Conclusion

FPVRGame is an egocentric vision environment with a hand pose interaction solution allowing the use of bare hands as a control for VR and Head Mounted Display scenarios, especially for low-end VR devices. Our scenario is focused on egocentric vision and presents a set of natural and comfortable hand poses, considering users' preferences.

The FPVRGame demonstrates a hand pose interaction solution, based on deep learning. Using a trained CNN model capable of recognizing hand poses, we recognized images from indoors and outdoors environments and without any illumination or background constraint. We achieved an average accuracy of 97.89%, which allowed smooth and comfortable human interaction through different usage scenarios.

We used a user-centered design approach to built and test the hand based 3D interaction vocabulary. Firstly, the questionnaire application was possible to select the initial vocabulary according to the users' preferences. Secondly, with the Wizard of Oz technique was possible to validate the preliminary vocabulary and also generate a large dataset (58,868 RGB images), which we made publicly available [29]. Finally, with the users' tests was possible to assess how natural the hand pose interaction was and compare it with traditional interaction devices using commercial low-end HMD's.

Chapter 7

Future Works

In future work, we intend to explore the weaknesses pointed out in the user feedback obtained during the evaluation of the FPVRGame, such as the fact that the user makes some gesture out of the camera angle and does not get any response from the system. This failure occurs due to the limitation of using only one input device, and we assume that adding the solution to a complementary input device using the concept of the Internet of Things (IoT) and associating it to the context of the game could solve the issue.

We can add to our solution the hand-pose segmentation features associated with a tracking system and estimating the real-world illumination. In a mixed reality-context, the usage of a more natural and immersive alternative to the game controllers, such as the user's hands, may drastically increase the game interface experience, allowing personalized visual feedback of the user's interactions [23].

References

[1] Al Maimani, A.; Roudaut, A. Frozen suit: Designing a changeable stiffness suit and its application to haptic games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 2440–2448.

[2] Alkemade, R.; Verbeek, F. J.; Lukosch, S. G. On the efficiency of a vr hand gesture-based interface for 3d object manipulations in conceptual design. *International Journal of Human–Computer Interaction 33*, 11 (2017), 882–901.

[3] Bowman, D. A.; McMahan, R. P.; Ragan, E. D. Questioning naturalism in 3d user interfaces. *Communications of the ACM 55*, 9 (2012), 78–88.

[4] Buxton, B. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.

[5] Cao, C.; Zhang, Y.; Wu, Y.; Lu, H.; Cheng, J. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3763–3771.

[6] Crankshaw, D.; Wang, X.; Zhou, G.; Franklin, M. J.; Gonzalez, J. E.; Stoica, I. Clipper: A low-latency online prediction serving system. In *NSDI* (2017), pp. 613–627.

[7] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), Ieee, pp. 248–255.

[8] Fushiki, T. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing 21*, 2 (2011), 137–146.

[9] Gavin, S. Flower ui: A gaze based interaction approach for mobile vr, Oct. 2017.

[10] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[11] Höll, M.; Oberweger, M.; Arth, C.; Lepetit, V. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces* (2018).

[12] Huang, G.; Liu, Z.; Weinberger, K. Q.; van der Maaten, L. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), vol. 1, p. 3.

[13] Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).

[14] JoliBrain. Deep detect, 2018. Accessed: 2018-09-20.

[15] Kanade, T.; Hebert, M. First-person vision. *Proceedings of the IEEE 100*, 8 (2012), 2442–2453.

[16] Kelley, J. F. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1983), ACM, pp. 193–196.

[17] Kim, M.; Lee, J.; Jeon, C.; Kim, J. A study on interaction of gaze pointer-based user interface in mobile virtual reality environment. *Symmetry 9* (2017), 189.

[18] Knierim, P.; Schwind, V.; Feit, A. M.; Nieuwenhuizen, F.; Henze, N. Physical keyboards in virtual reality: Analysis of typing performance and effects of avatar hands. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 345:1–345:9.

[19] Koller, O.; Ney, H.; Bowden, R. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, USA, June 2016), pp. 3793–3802.

[20] Lee, S.; Park, K.; Lee, J.; Kim, K. User study of vr basic controller and data glove as hand gesture inputs in vr games. In *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)* (June 2017), pp. 1–3.

[21] Li, Y.; Ye, Z.; Rehg, J. M. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 287–295.

[22] Lin, J.; Harris-Adamson, C.; Rempel, D. The design of hand gestures for selecting virtual objects. *International Journal of Human–Computer Interaction* (2019), 1–7.

[23] Marques, B. A. D.; Clua, E. W. G.; Vasconcelos, C. N. Deep spherical harmonics light probe estimator for mixed reality games. *Computers & Graphics 76* (2018), 96–106.

[24] Molchanov, P.; Yang, X.; Gupta, S.; Kim, K.; Tyree, S.; Kautz, J. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4207–4215.

[25] Mortensen, D. Natural user interfaces–what are they and how do you design user interfaces that feel natural. *The Interaction Design Foundation* (2017).

[26] NVIDIA. NVIDIA inference server, 2018. Accessed: 2018-09-20.

[27] NVIDIA. NVIDIA tensorrt, 2018. Accessed: 2018-09-20.

[28] Oculus. Oculus homepage., 2018. Accessed: 2018-09-10.

[29] Oliveira, E. Dataset from egocentrics images for hand poses recognition., 2018. Accessed: 2018-09-10.

[30] Pereira, A.; Wachs, J. P.; Park, K.; Rempel, D. A user-developed 3-d hand gesture set for human–computer interaction. *Human factors 57*, 4 (2015), 607–621.

[31] Pirker, J.; Pojer, M.; Holzinger, A.; Gütl, C. Gesture-based interactions in video games with the leap motion controller. In *International Conference on Human-Computer Interaction* (2017), Springer, pp. 620–633.

[32] Proença, P. F.; Gao, Y. Splode: Semi-probabilistic point and line odometry with depth estimation from rgb-d camera motion. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on* (2017), IEEE, pp. 1594–1601.

[33] Proença, P. F.; Gao, Y. Probabilistic rgb-d odometry based on points, lines and planes under depth uncertainty. *Robotics and Autonomous Systems 104* (2018), 25–39.

[34] Rautaray, S. S.; Agrawal, A. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review 43*, 1 (Jan 2015), 1–54.

[35] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*, 3 (2015), 211–252.

[36] Sagayam, K. M.; Hemanth, D. J. Hand posture and gesture recognition techniques for virtual reality applications: a survey. *Virtual Reality 21*, 2 (Jun 2017), 91–107.

[37] Sampson, H.; Kelly, D.; Wünsche, B. C.; Amor, R. A hand gesture set for navigating and interacting with 3d virtual environments. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)* (2018), IEEE, pp. 1–6.

[38] Schmid, K.; Hirschmüller, H. Stereo vision and imu based real-time ego-motion and depth image computation on a handheld device. In *2013 IEEE International Conference on Robotics and Automation* (2013), IEEE, pp. 4671–4678.

[39] Son, Y.-J.; Choi, O. Image-based hand pose classification using faster r-cnn. In *Control, Automation and Systems (ICCAS), 2017 17th International Conference on* (2017), IEEE, pp. 1569–1573.

[40] Sra, M.; Xu, X.; Maes, P. Breathvr: Leveraging breathing as a directly controlled interface for virtual reality games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 340:1–340:12.

[41] Sutherland, I. E. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I* (1968), ACM, pp. 757–764.

[42] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[43] Tewari, A.; Grandidier, F.; Taetz, B.; Stricker, D. Adding model constraints to cnn for top view hand pose recognition in range images. In *ICPRAM* (2016), pp. 170–177.

[44] VIVE™. Vive™ | discover virtual reality beyond imagination, 2018. Accessed: 2018-09-10.

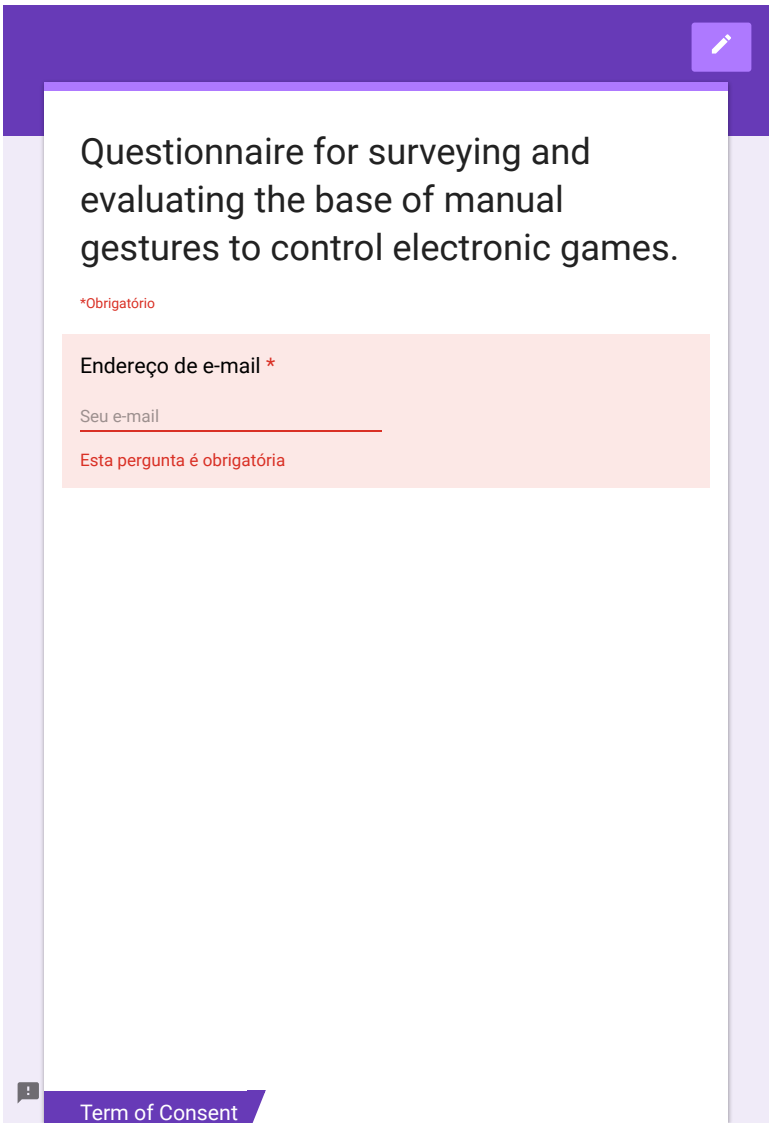
[45] Wang, H.; Schmid, C. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 3551–3558.

[46] Yousefi, S.; Kidane, M.; Delgado, Y.; Chana, J.; Reski, N. 3d gesture-based interaction for immersive experience in mobile vr. In *Pattern Recognition (ICPR), 2016 23rd International Conference on* (2016), IEEE, pp. 2121–2126.

[47] Zhang, C.; Xue, Q.; Waghmare, A.; Meng, R.; Jain, S.; Han, Y.; Li, X.; Cunefare, K.; Ploetz, T.; Starner, T.; Inan, O.; Abowd, G. D. Fingerping: Recognizing fine-grained hand poses using active acoustic on-body sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 437:1–437:10.

[48] Zhang, Y.; Cao, C.; Cheng, J.; Lu, H. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia 20*, 5 (2018), 1038–1050.

APPENDIX A – Questionnaire for Surveying and Evaluating the Base of Manual Gestures to Control Electronic Games



Research Title: Interacting in Virtual Reality games using recognition of hand poses captured from the player's point of view.

Researchers in charge: Paulo Roberto Possas (UFF), Eder de Oliveira (Master's degree - UFF), Professor Luciana Cardoso de Castro Salgado (Counselor - UFF), Professor Daniela G. Trevisan (Counselor - UFF), Professor Dr. Esteban W. G. Clua (Advisor - UFF), Professor Cristina Nader Vasconcelos (Counselor - UFF).

Dear volunteer,

You are being invited as a volunteer to participate in the research "Interacting in Virtual Reality games using recognition of hand poses captured from the player's point of view." In this study we intend to raise and evaluate different types of manual gestures to control the characters of a game.

FREE AND CLOSED CONSENT TERM FOR SCIENTIFIC RESEARCH PARTICIPATION.

For this study we will adopt the following procedure (s): Respond to a questionnaire (online) whose objective is to raise and evaluate a base of standard hand poses for the implementation of a tool that is intuitive and comfortable.

To participate in this study you will have no cost, nor receive any financial advantage. You will be free to attend or refuse to attend. Your participation is voluntary and refusal to participate can be done at any time. The researcher will treat your identity with professional secrecy standards. You will not be identified in any publication that may result from this study.

The search results will be at your disposal when finalized. Your name or material indicating your participation will not be released without your permission. The data and instruments used in the research will be archived with the researcher responsible for a period of 5 years, and after that time will be destroyed.

To obtain the material of your participation, or for any other information regarding the questionnaire, please contact:
Paulo Roberto Possas
Address: Rua São Pedro de Itaipu, 316
CEP: 24355-220 / Niterói – RJ
E-mail: paulopossas@id.uff.br

or

Eder de Oliveira
Email: eder.oliveira@ifmt.edu.br

Professional address for student location:
Av. Gal. Milton Tavares de Souza, s/n - São Domingos, Niterói - RJ, 24210-310.

I agree to everything that has been previously quoted and freely

give my consent by submitting this completed form. *

- ◯ Yes.
- ◯ No.

Esta pergunta é obrigatória

Página 1 de 5

PRÓXIMA

Nunca envie senhas pelo Formulários Google.

Questionnaire for surveying and evaluating the base of manual gestures to control electronic games.

*Obrigatório

First, help us set up your user profile.

What is your name? *

Sua resposta

Whats is your nationality? *

Escolher

What is your age range? *

- ( ) Less than 18
- ( ) 18 - 24 years
- ( ) 25 - 30 years
- ( ) 30 - 35 years
- ( ) More than 35 years

Do you have color blindness? *

- [ ] No
- [ ] Yes

If you have any problems that limits or makes it impossible to move either or both hands, please specify.

Sua resposta

Evaluate how much you have heard of the technologies below from 1 to 5. *

Your knowledge of technology corresponds to how much you have read and / or studied about technology in books, magazines, or any other form of media. Admit that 1 represents no knowledge and 5 represents a lot of knowledge.

	1	2	3	4	5
Games for computer or consoles (XBOX ONE, PlayStation 4, Nintendo Wii U, etc.)	○	○	○	○	○
Mobile games (Android, iOS, Windows Phone, etc.)	○	○	○	○	○
Virtual Reality Games with Head Mounted Displays (HTC VIVE, Oculus Rift, PlayStation VR, Microsoft Hololens, etc.) for computers or consoles	○	○	○	○	○
Virtual Reality Games with Head Mounted Displays (Google Daydream, Cardboards in general) for mobile phones	○	○	○	○	○
Gesture recognition tools (Kinect, Leap Motion, Myo, etc.)	○	○	○	○	○

Rate your familiarity with technologies below from 1 to 5. *

Your familiarity with the technology corresponds to how much you have used or interacted with that technology in your life. Admit that 1 represents never having used the technology and 5 represents using the technology on a daily basis.

	1	2	3	4	5
Mobile games (Android, iOS, Windows Phone, etc.)	○	○	○	○	○
Virtual Reality Games with Head Mounted Displays (HTC VIVE, Oculus Rift, PlayStation VR, Microsoft Hololens, etc.) for computers or consoles	○	○	○	○	○
Virtual Reality Games with Head Mounted Displays (Google Daydream, Cardboards in general) for mobile phones	○	○	○	○	○
Gesture recognition tools (Kinect, Leap Motion, Myo, etc.)	○	○	○	○	○

What is your dominant hand? If ambidextrous, check the option that you use the most every day. *

Your answer will change the next questions you will answer.

- ○ Right
- ○ Left

Questionnaire for surveying and evaluating the base of manual gestures to control electronic games.

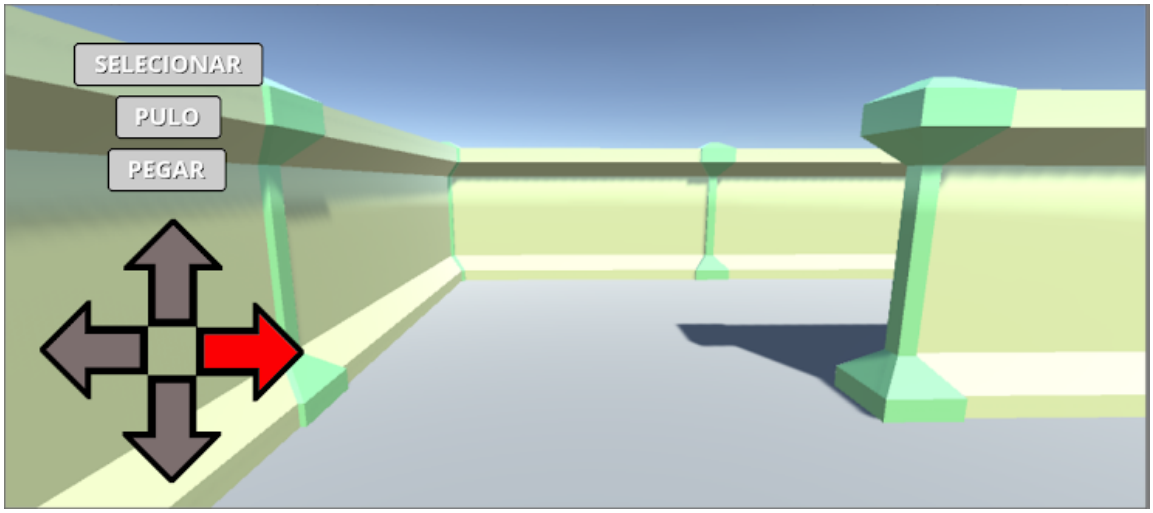
*Obrigatório

Now imagine that you are using a virtual reality glasses similar to the image below, which is able to recognize the gestures that you perform with your hands. As you said earlier that your dominant hand was right, the images shown in this section will all be right handed.

In this section we will present images that represent situations of decision making in a game. In each of the images we will specify the command to be performed (it will be indicated in red) and you should choose the image with the manual gesture that is more natural and comfortable. You can choose only one option for each question. If you do not like any of the moves, please specify the movement you would use.

Which of the following gestures would you choose to represent

the "turn right" command? *



1

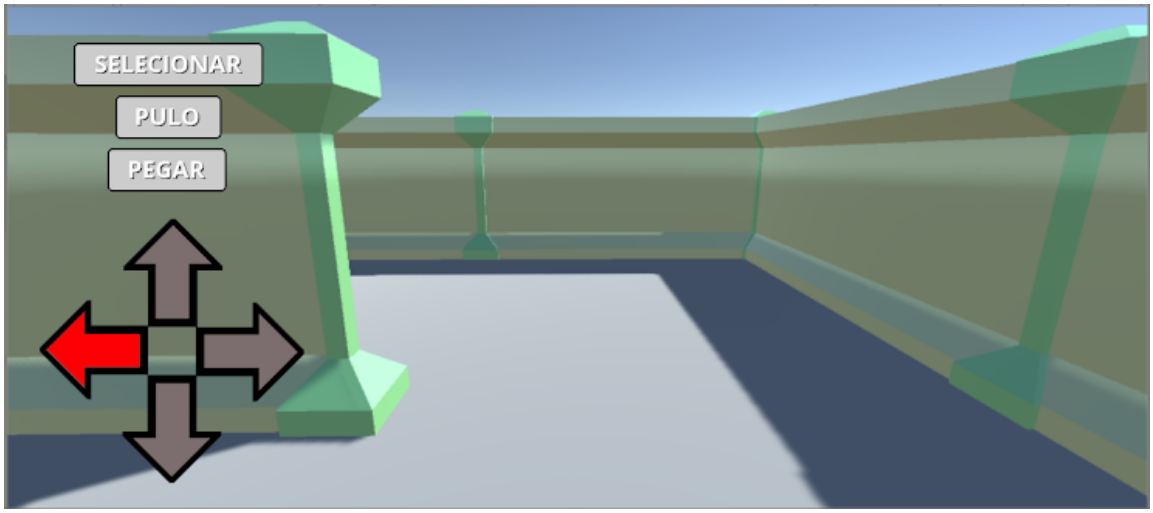
2

Outro:

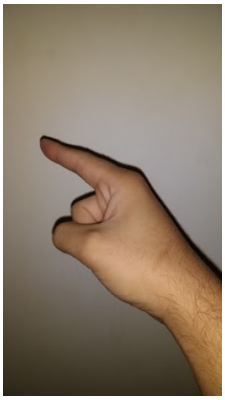
3

Which of the following gestures would you choose to represent

the "turn left"command? *



1



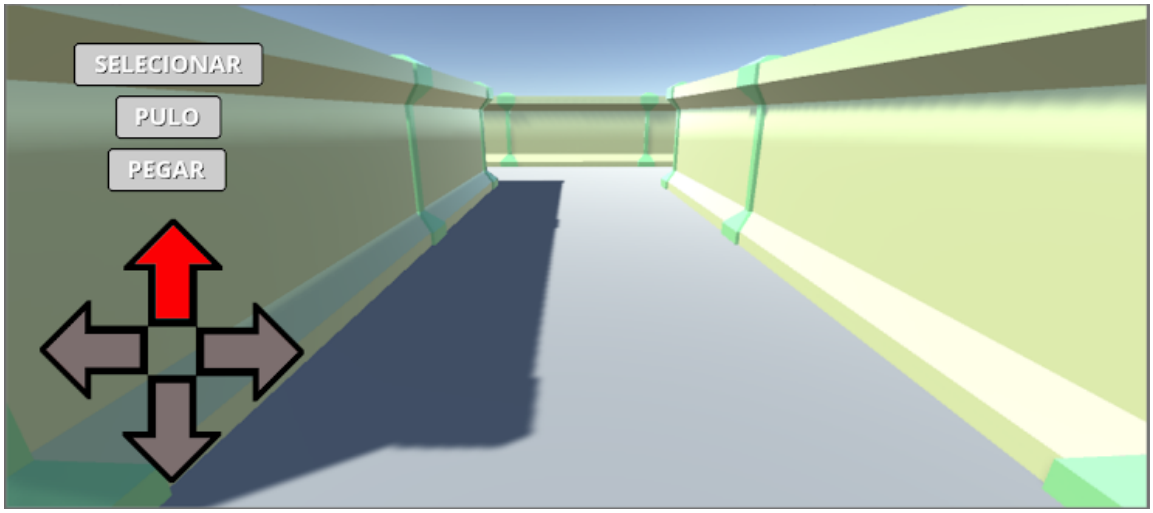
2

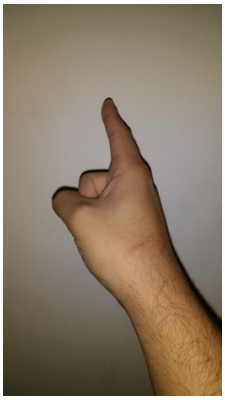
Outro:

3

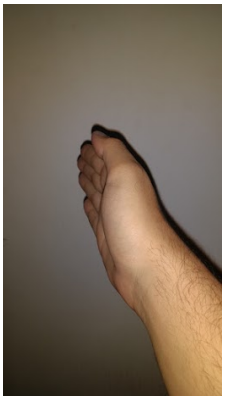
Which of the following gestures would you choose to represent

the "move forward" command? *



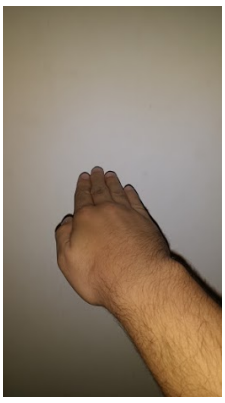


1



2

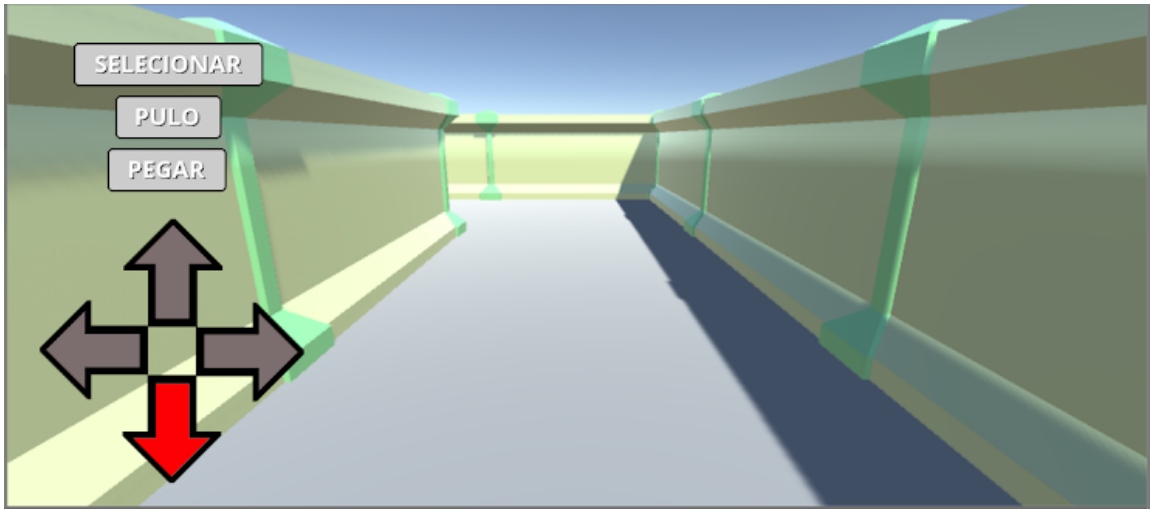
Outro:

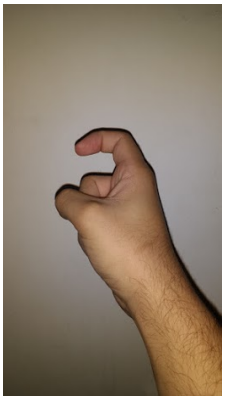


3

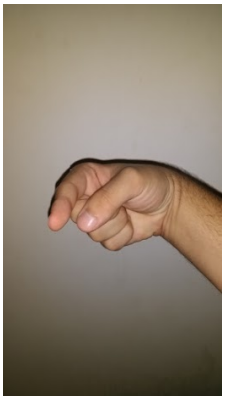
Which of the following gestures would you choose to represent

the "move backwards" command? *



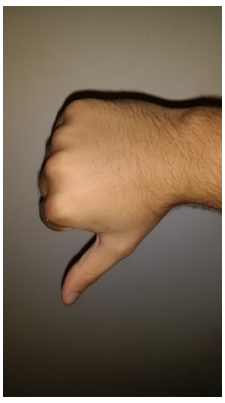


1



2

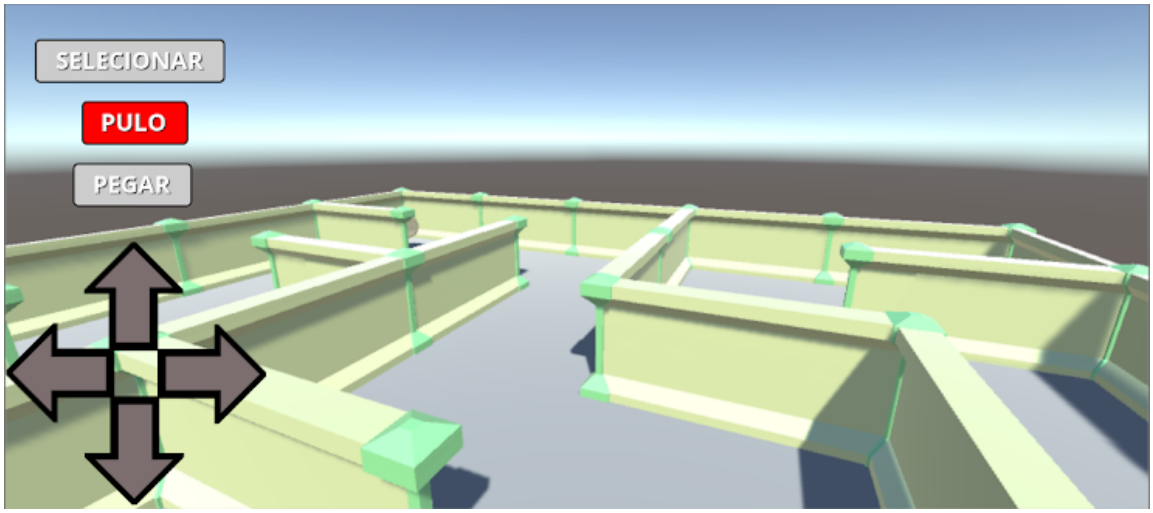
Outro:

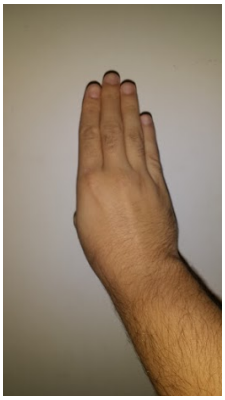


3

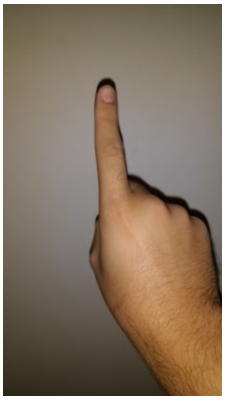
Which of the following gestures you choose to represent the

"jump" command? *



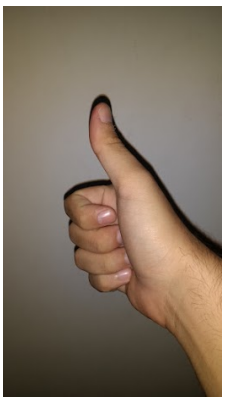


1



2

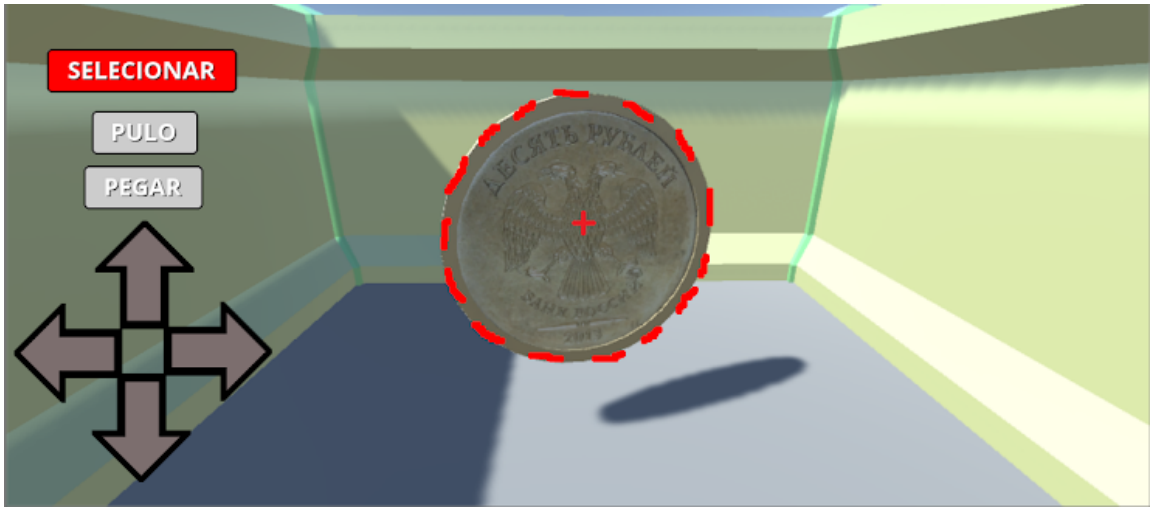
Outro:



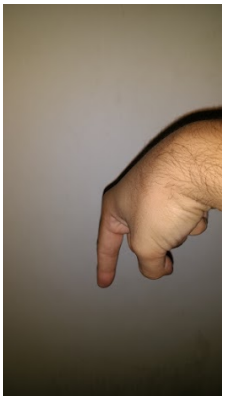
3

Which of the following gestures would you choose to represent

the "select coin" command? *

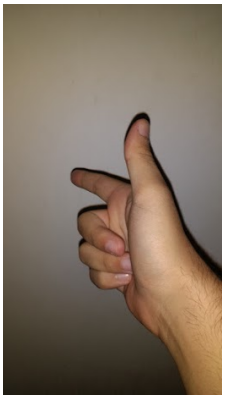


1



2

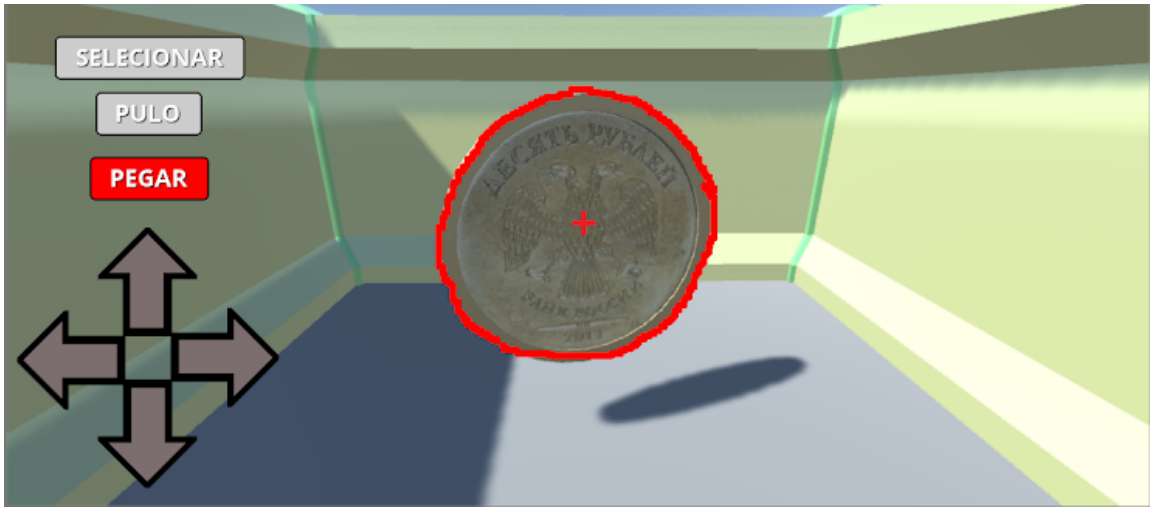
Outro:

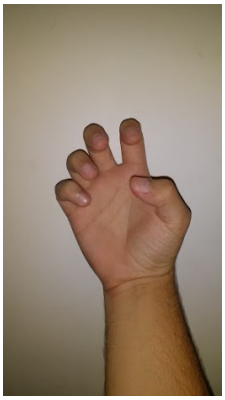


3

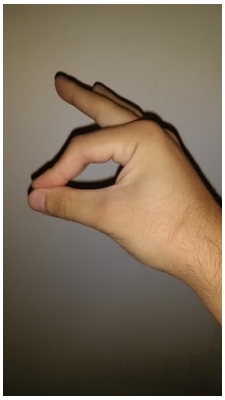
Which of the following gestures would you choose to represent

the "grab the coin" command? *





1



2

Outro:

3

VOLTAR PRÓXIMA

Questionnaire for surveying and evaluating the base of manual gestures to control electronic games.

Comments

Please leave here your critics and comments about the questionnaire and / or research. Your feedback is very welcome. Thank you!

Sua resposta

Envie-me uma cópia das minhas respostas.

Página 5 de 5

VOLTAR ENVIAR

Nunca envie senhas pelo Formulários Google.