TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Oleksandra Zamana 184083IVSB

Personal identification in Virtual and Augmented Reality: Control Systems Perspective

Bachelor's thesis

Supervisors: Aleksei Tepljakov, PhD
Saleh Al-saleh, MSc

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Oleksandra Zamana 184083IVSB

Isikutuvastus virtuaalreaalsuses: juhtimissüsteemide vaade

Bakalaureusetöö

Juhendaja: Aleksei Tepljakov, PhD

Saleh Al-saleh, MSc

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Oleksandra Zamana

Abstract

Virtual and Augmented reality is a rapidly growing, thriving domain. While a lot of benefits come from using VR/AR, involved technology poses significant privacy and security risks. Even biometric data collected during trivial experiments, where subjects perform repetitive tasks, might be used for unique human identification, which makes it a candidate for ethical and legal discussions. The research question of this thesis is if we can consider data collected during trivial VR experiences as personal data: this thesis strives to establish if personal identification in a Virtual Reality environment using data collected during trivial experiments is possible and thus state if such data is sensitive. In other words, if it can be used to identify a human being.

Even though a number of researches have been conducted on the topic of human identification in VR/AR, experiments were designed to enhance the possibility of such identification. On the contrary, this work is dealing with natural, real-life behaviour of subjects interacting with so called digital twins of control objects control systems in Virtual Reality in the area of automatic control systems design.

Research was conducted in such stages: experiment preparation, experiment and data collection, data pre-processing, data analysis, gathering and presenting the results. 37 subjects took part in the experiment, out of which: 17 identify as female, 20 identify as male; 33 are right-handed, 4 are left-handed. Approximately 40% of the participants had experience in VR/AR in the past. Hypothesis is the thesis is that data collected during a trivial experiment can show solid identification accuracy among a group of participants. This thesis compares results given by five machine learning classifiers and 40 different variations of features and metrics, achieving the accuracy of 75% success rate of the identification for the group of 32 people.

This thesis is written in English and is 47 pages long, including 7 chapters, 14 figures and 2 tables.

Annotatsioon

Virtuaal- ja liitreaalsus on kiiresti kasvav, kiirelt arenev valdkond. Kuigi virtuaal- ja liitreaalsuse kasutamine pakub palju eeliseid, kujutavad sellega seotud tehnoloogiad endast märkimisväärseid privaatsus- ja turvariske. Isegi lihtsates katsetes, milles katsealused täidavad korduvaid ülesandeid, võib tegevuste käigus kogutud biomeetrilisi andmeid kasutada inimese ainulaadseks identifitseerimiseks, andes sellega alust eetilisteks ja õiguslikeks aruteludeks. Lõputöö eesmärgiks on välja selgitada, kas isikupärastamine tehisreaalsuse keskkonnas, lihtsate katsete käigus kogutud andmete abil, on võimalik ja kas sellised andmed on tundlikud, teisisõnu, kas neid saab kasutada inimese tuvastamiseks. Ehkki virtuaal ja liitreaaluses inimeste tuvastamise teemal läbi viidud mitmeid uuringuid, olid katsed kavandatud sellisel moel, mis suurendas identifitseerimise võimalust. Seevastu see töö käsitleb katsealuste loomulikku interaktsiooni nn. digitaalse kaksikuga virtuaalreaalsuses automaatjuhtimissüsteemide arendamise valdkonnas.

Uuring viidi läbi järgnevates etappides: katse ettevalmistamine, katse ja andmete kogumine, andmete eeltöötlus, andmete analüüs, tulemuste kogumine ja esitamine. Katses osales 37 inimest, kellest 17 määratles end naisena, 20 määratles end mehena, 33 olid paremakäelised ja neli vasakukäelised. Ligikaudu 40 protsendil osalejatest oli varasem VR / AR-i kasutamiskogemus. Selles lõputöös võrreldakse viit masinõppe klassifitseerijat ning 40 erineva tunnuste ja mõõdikute variatsiooni tulemusi, saavutades 75-protsendilise identifitseerimise täpsuse 32-liikmelise grupi jaoks.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 47 leheküljel, 7 peatükki, 14 joonist, 2 tabelit.

List of abbreviations and terms

VR	Virtual Reality
AR	Augmented Reality
XR	Extended Reality
GDPR	General Data Protection Regulation
SUS	System Usability Scale
CSV	Comma-Separated Values
ML	Machine Learning
DP	Data Preprocessing
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
MDT	Missing Data Treatment

Table of contents

List of figures

List of tables

1 Introduction

In the context of this thesis four interconnected topics will be covered: Virtual and Augmented reality, personal data, control systems and machine learning.

1.1 Virtual and Augmented Reality

Virtual and Augmented reality experience is all about immersing the user into digital space created programmatically using computer modelling and accessed through devices as headsets, controllers or glasses. Both Virtual and Augmented reality have been present in the scientific arena for a relatively short amount of time. Despite it, this area of technology is expanding on an everyday basis: it is eagerly explored by both business and science worlds. Forecasts regarding market value differ from $95 billion [1] to $569 billion market in 2025 [2]. At the present time, global virtual reality market size was valued at USD 15.81 billion in 2020 [3]. As it stands, 26 million of these devices are currently owned by private consumers [9]. According to the latest estimates, as many as 5.5 million units are set to be further shipped to customers worldwide in 2020 [10]. As we can tell from the numbers, VR and AR are attracting attention and growth of this technology is not expected to lower its pace in the foreseeable future.

1.2 Personal Data

Virtual and Augmented reality brings lots of opportunities into our daily lives through its influence on manufacturing, healthcare, energy, retail, and training and development [4], but as any technology meant for close interaction with humans, Virtual and Augmented reality poses significant privacy and security risks.

From the GDPR standpoint personal data is considered to be are any information that is related to an identified or identifiable natural person. The data subjects are identifiable if they can be directly or indirectly identified. At its most basic form, whenever you differentiate one individual from others, you are identifying that individual. Thus, any individual who can be distinguished from others is considered identifiable [8]. During the experience in the VR/AR participants go through a set of different activities, usually repetitive ones. At the time when artificial intelligence is progressing and therefore machine learning algorithms are getting more and more advanced, the repetitiveness makes this kind of data a promising material for personal identification, even though it is not as obvious as fingerprints or face structure [6]. Ethical concerns regarding data protection have been raised previously and are likely to become even more acute.

1.3 Research Question and Background

Now that it is clear that any data that can be used to distinguish one individual from another is considered to be personal from the GDPR perspective, the question of this thesis is if we can consider data collected during trivial VR experiences as such. As a trivial experiment in this research an educational experiment where subjects get to know Control Systems, in particular a 3D crane, is brought up. The research group was occupied by the problem of teaching the Control Systems subject more interactively in order to improve students' interest. Hands-on experiments with control objects are immensely helpful in teaching system theory and control system design, thus, as a proof of concept, a digital twin of a 3D crane was created, and students could interact with this twin through a VR application [27].

It is important to mention, that during data processing such common practices as filtering out first runs of the experiment to avoid learning errors were not implemented in order to achieve as natural experiment as possible. Design of the experiment was solely inspired by the needs of the research, assessing the feasibility of VR usage for educational purposes, and not by the improving identification results: users did not have any strict time frame or guidance for movement, participants were free to exhibit any behavior as long as they could give feedback on the usage of the digital twin. This factor makes the research unique, as it places participants in a situation much more similar to the real life. It is also crucial to understand that user identification is not the same as user authentication. The goal of identification is to identify a user from among a group of known users, while authentication focuses on independently identifying a user from another user who is likely to use the system. This thesis is focusing on the problem of identification [12].

This work is the results of the combined efforts of the author and a research group based in the Tallinn University of Technology. Whenever the term "research group" is used, it means that action was performed without the author's direct participation. Participants of the experiments will be referred to as "users", "subjects" or "participants" interchangeably. The following Section 4 "Experimentation" covers experiment design.

2 Literature Review

The following section gives an overview of the state of art and summarizes studies that inspired this research and were taken as a reference through the whole process of writing this thesis, as well as covering the existing legal and ethical concerns about data collection in VR/AR.

2.1 Related Work

Reviewing the materials present online it is possible to conclude that earlier several pieces of research were conducted that prove that human identification manipulating the data collected from VR/AR equipment is possible and, as a matter of fact, it is presenting incredible results. Some experiments showed the results up to 92.85% accuracy [7]. Others, like analysis of the approach using blink and head-movement data, showed the accuracy of 94% [12]. It has to be highlighted that none of the researches conducted earlier has a motivation to work with data collected during the trivial experiments and were designed with an intent of further user identification.

Related work provides us with examples of the experiment designs, set of features and machine learning algorithms used. Research “Behavioral Biometrics in VR” strives to solve such problems as continuous authentication in the VR experiences. It is inspired by the idea that every person moves uniquely, thus using the movement data it is possible to identify the subjects. It presents a study of 22 people performing VR tasks such as pointing, grabbing, typing and walking. In the research data is collected from the headset and controllers, and additionally eye tracking is enabled for monitoring eye movement. Experiments were conducted in two sessions: during the first session data was collected for training, and in the second – for testing. The first two runs and first two targets were filtered to avoid learning effects. Research uses Random Forest and Support Vector Machine classifiers from sklearn Python library.

Results of the research were taken into consideration while selecting features, as they showed that the best body motion to identify and authenticate users are head motions and distances between the devices. Also, it was stated that pointing gives more accurate results than grabbing due to more diverse movements.

As this topic has already been brought up in multiple research works there is already quite a stable foundation for this thesis: scientists have already tried a certain number of methods and approaches to achieve results. Of course, as these devices have not entered our lives on a sufficient level yet, there are still a lot of details left yet to discover and to study, for example:

- further evaluation of classifiers, features, and variations and combinations;
- investigation of other treatments of the temporal aspect of the data, including, for example, extracting features for rolling windows over time instead of aggregation per action;
- investigation of the frequency domain of features instead of the time domain;
- investigation of potential improvements via aggregation of evidence over time (for example, decision after observing multiple actions) [5].

Current research provides further evaluation of classifiers, features, variations and combinations, and the evaluation of multiple actions.

2.2 Security and Privacy Concerns

While results and success rate of personal identification in VR were described in the previous section, the following section talks about what privacy and security concerns arise from collecting this data.

Authors Kent Bye, Diane Hosfelt, Sam Chase, Matt Miesnieks, Taylor Beck of the study “The ethical and privacy implications of mixed reality” [15] believe that biometric information presents particularly difficult problems because there is no way to retrieve or change such information once exposed. Also, it provides methods for fingerprinting users by their physical attributes. There is a concern that some companies are already using VR during interviews to determine how applicants will react to various scenarios, which creates a threat for the privacy of applicants. The question is if it is ethical to use all the wide range of data that VR can collect and interpret. It is known that biometrics provide insight into such aspects as involuntary nonverbal reactions, which can be used to rate users' opinions of experiences, but this same data can also be used to reveal very sensitive facts about a user, such as diseases or sexual preferences. Authors believe that we need a new ethical design principles to help guide our experiential design frameworks, as even if we are aware that our data is being collected, we do not know how sensitive it is.

Ellysse Dick in their study "Balancing User Privacy and Innovation in Augmented and Virtual Reality" [16] believes that we need to bring these concerns not only to the ethical level but also to the legal one. It highlights that information collected in VR/AR is different from what most other consumer technology devices and is much more sensitive. The author states that there are unique risks of aggregating sensitive information and thus there is a need to adopt mitigation measures that were designed for other consumer technologies into VR/AR by changing the laws accordingly. "Government agencies and industry should develop voluntary guidelines for AR/VR developers to secure users' privacy through transparency and disclosure practices, user privacy controls (including opt-out mechanics), information security standards, and considerations for the unique risks presented by biometric identifying and biometrically derived data", — it states.

In the context of current work, we are talking about two types of data that the study categorized:

- observed data: information an individual provides or generates, which third parties can observe but not replicate, such as heart rate and coordinates of body parts;
- computed data: new information AR/VR technologies infer by manipulating observable and observed data, such as biometric identification.

Privacy considerations for both types are similar: user anonymity and autonomy, security of sensitive provided information, the potential for discriminatory use of such information. Even though biometric data is necessary for reconstructing physical experiences in fully virtual spaces, users need to understand how and why various AR/VR services collect and share their data, so that they can make informed decisions about the information they choose to share. Moreover, users should understand what information can be inferred from the data they provide, as disclosure of such information can lead to significant reputational harm or embarrassment when the nature of the inferred information is particularly sensitive or highly personal.

2.3 Virtual and Augmented Reality and Law

In their work “Research Handbook on the Law of Virtual and Augmented Reality” [24] authors Woodrow Barfield and Marc Jonathan Blitz are talking about the ways to apply existing legislation to Virtual Reality problems and disputes that arose on this ground. Courts are already struggling with cases that fall on the edge of real and virtual worlds, such as the following examples: Pokemon Go users and creators are blamed for trespassing onto private land and causing damage. As another example, Linden Lab was sued by an individual whom it barred from future access to its virtual world, while this individual spent lots of time and money creating their virtual property. Some people considered themselves to be sexually assaulted when their characters in the virtual world were touched without consent.

In the article “Virtual reality: top data protection issues to consider” [26] authors are referencing GDPR article 4(14), saying that data collected in VR (head location, heartbeat, etc) is identifiable as biometric data under article 4(14) of the GDPR, “personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopy data”. GDPR provides that the processing of biometric data (to uniquely identify a natural person) shall be prohibited unless the data subject has given explicit consent to the processing.

Even though GDPR covers the question of data collected during the experience in Virtual Reality, Sheri B. Pan in work “Get to know me: protecting privacy and autonomy under big data's penetrating gaze” [25] states that academic, legal, and industry conceptions of informational privacy in the past have failed to consider the harm potentially posed by big data's capability of inferring new personal information. It is difficult for a person to fully realize what personal information is being collected for analysis and cannot predict what personal information such data will infer about them. As a result, people unknowingly share information they neither intended nor wanted to reveal. As a fundamental theory of privacy defines privacy as the control over personal information, the study questions how the control theory evaluates privacy where some personal data, voluntarily collected from the person, can be analyzed to infer other information never disclosed by the person.

Thus, in connection with this thesis, it must be clear for users that biometric data collected during a trivial experience can be used to identify them with a soft biometric accuracy level, and while currently we are talking about relatively small groups, in the future results can drastically improve as identification tools are evolving and data is becoming more and more valuable for businesses.

2.4 The Gap in the Literature

As a researcher of the topic of personal identification in Virtual and Augmented reality, author felt that there is a lack of thorough description of data manipulations from the beginning of the process to the end, which makes it harder to adopt best practices of data preprocessing and processing. Especially it can be noticed in regards to data preprocessing: for example, a lot of works did not specify classifiers or feature selection process.

As was stated before, studies analyzed during literature review do not include real-life experiments, where subjects are free to exhibit natural behavior outside of strict boundaries created to ensure better results of data processing.

This work aims to close the gap by analyzing data collected during the trivial experiment and providing a detailed step by step descriptions of data handling, from the data collection to gathering the results, including details such as parameters used for classifiers.

3 Methodology

In the present thesis, an empirical study is preferred to a theoretical one, as the hypothesis chosen creates room for experimenting, observation, and further analysis. To obtain the data needed to conduct the analysis and provide conclusions a number of the documented experiments need to take place. Further, you will find the list of the methodologies used in the research work:

1. Documentary analysis (literature review);
2. Statistics (preprocessing);
3. Machine learning:
    1. Logistic Regression classifier;
    2. Support Vector classifier;
    3. Random Forest classifier;
    4. KNN classifier;
    5. Decision Tree classifier;
4. Presentation of data:

    Data gathered from the documentation can be presented in various forms, like line plots and bar plots, as the topic is heavily statistical.

5. Comparison and analysis:

    Comparison and analysis are two important methods to use in the research work in question, as in the end of the work in question it will be possible to compare different methods used in human identification and analyze results obtained. Furthermore, impact of the information retrieved from the data can be estimated.

The timeline of the research is presented by the chart (Figure 1).

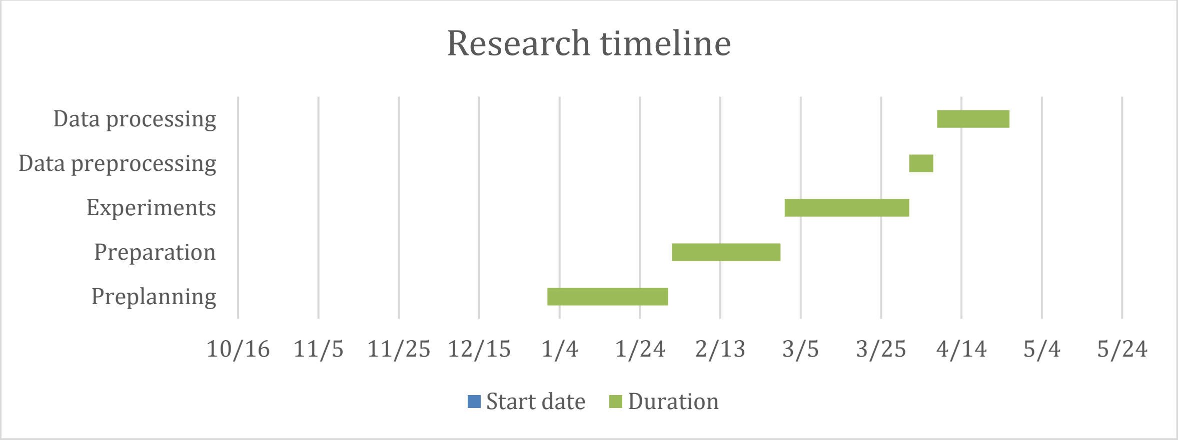

Figure 1. Research Timeline

4 Experimentation

4.1 Experiment Design

This section talks in detail about the essence and structure of the conducted experiments.

As was stated earlier, the experiment design was not meant to enhance the identification results, thus subjects were performing common tasks in a free manner. The experiment’s primary goal was to estimate if experience in VR makes it easier for participants to visualize and comprehend some concepts from the Control Systems subject from the university program. Control systems subject studies systems, which provide the desired response by controlling the output. In the experiment, the focus was on interaction with a 3D crane. The crane is called a “3D crane” because it has three degrees of freedom – it can move the payload along three different axes. While a 3D crane is moving the load, the load is going to be naturally swinging. The idea of the experiment is to teach subjects about swing compensation – a controller that can be applied through the movement of the crane to minimize a swing.

4.1.1 Preparation

In the preparation stage, the team agreed on the equipment that was going to be used – a Vario headset was used along with HTC motion controllers and HTC trackers. The headset acts as an interface: participants receive a visual picture from a headset. Controllers are used for interaction with any actors in the VR if needed. Trackers are responsible for the collection of data from the controllers and a headset to make sure the experience is smooth. Using these tools it is possible to collect time-stamped information about the location of hands and head during the experiment, along with information about eye movement, and after this process data to extract some valuable information from it.

The preparation stage took place in February 2021. First of all, a registration form to find subjects was created. Most of the subjects are current university students or alumni. During this time an Informed Data Usage consent form and System Usability Form were created. An informational presentation, simulation in VR and

MATLAB/Simulink environments were finalized by the research group. After, list of participants was created and split into time slots. The estimated average time of one experiment is 45 minutes and the final number of participants is 37. Experiments took place during four weekends in March 2021.

Before the subject arrives at the laboratory where experiments were conducted, the following had to be done or ensured to be done:

- Prepare the area:
    - Ensure participant safety by cleaning and disinfecting surfaces that the previous participant comes into contact with;
    - Change ID value on the whiteboard.
- Prepare MATLAB/Simulink:
    - Restart previously opened the program;
    - Set all the values to default.
- Prepare the VR equipment
    - Charge headset and controllers;
    - Place headset and controllers on starting positions;
    - Make sure that trackers are places correctly and there are no obstacles between trackers and the area of the experiment.
- Reopen the SUS form.

4.1.2 Stage 1

As the goal of the experiment is educational, participants had very little background on the topic. Once the subject arrived in the laboratory room, they had a chance to read consent forms, sign them and read through the informational presentation about Control Systems basics and a 3D crane (Figure 2). After, they proceeded to the first part of the experiment, implemented in the MATLAB/Simulink environment. MATLAB is a de-facto standard platform for designing control systems. Simulink is a visual editor that allows to build systems and controllers intuitively. During this part of the experiment, the subject was trying to control or interact with a 3D crane through the program using desktop.

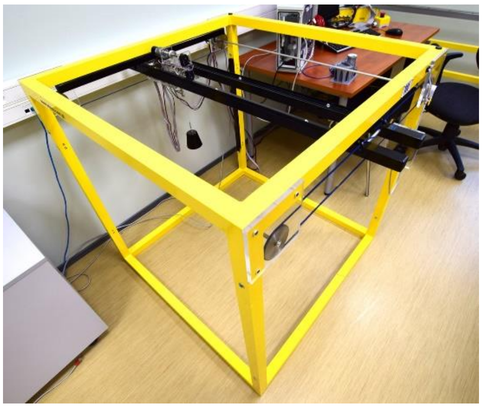

Figure 2. Real-life 3D Crane

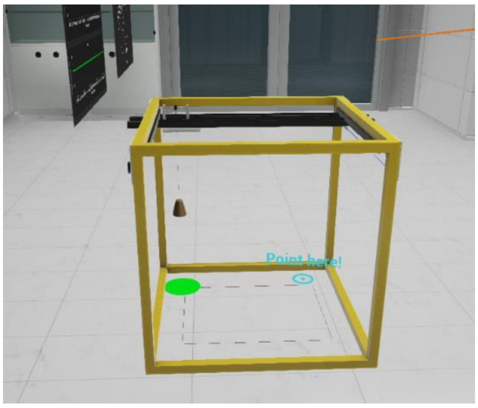

Figure 3. Digital Twin of the 3D Crane

4.1.3 Stage 2

Once the first part of the experiment is over, the subject proceeds to the second part of the experiment, where they interact with the model of the 3D crane in Extended Reality (XR) (Figure 3). Application for the experience is written using Unreal Engine, a game engine developed by Epic Games, but before that, subjects receive a short explanation about Virtual Reality headset and controllers. The last preparation step was to complete eye calibration – a set of actions that allows the headset to capture eyesight correctly and adjust the visual so that it appears clear.

Once everything is set, participants find themselves in the simulated virtual environment: in our case in the laboratory room. They spot the marking on the ground “Walk here” (Figure 4) and proceed there. From that spot, they will be able to see a virtual 3D crane model (Figure 3). The goal is to point on the ground under the 3D crane using a controller and thus make the crane move load from the arbitrary position A to the arbitrary position B. At some point, the subject turns on swing compensation and observes the results. During the experiment subjects can see the graphs with flowing data about angles – they can try to grab the graph and position it wherever they want in the space around them.

Figure 4. Virtual Laboratory

The main objective of this stage is that participants understand how the crane behaves with a swing compensation turned on and off. As long as this objective is achieved, for the rest of the experiment subject is free to explore the 3D crane model or graphs, walk within the limits of the area provided. After every experiment data collected during the experiment was uploaded to the private cloud repository in the ZIP format, password protected.

4.1.4 Feedback

After the second stage is completed, subjects were asked to fill a SUS form to give feedback. Feedback helps to evaluate what system is more convenient and understandable for usage: MATLAB/Simulink or a VR/AR implementation of a 3D crane. Feedback was used in the paper by the research group.

4.2 Data Collection

The following section describes the process of data collection during the experiments and its storage. It is also describing what kind of data was collected.

Data from 37 male and female subjects was captured ranging from 18 to 34 years old, out of which: 17 identify as female, 20 identify as male; 4 participants have left hand as their dominant hand, 33 are right-handed. 89% of the participants are university students, alumni or staff. Around 40% of subjects had experience in Virtual Reality before.

As involving behavioral biometrics is one of the most promising methodologies to uniquely identify the user, it was one of the inspirations of this work [5]. The best body motions to identify users from a range of available actions are head motions, and distances between the devices (controllers) [5]. Activities for the subjects to go through were selected based on the specifics of the educational experiment. From the perspective of data collection, we are interested in the second part of the experiment, where participants use a headset and controllers and perform certain actions in VR. From what is described in the Section 4.1.3 it is possible to identify three main activities performed in the experiment in question:

- walking from the starting position to the target;
- pointing on the ground to move the load;
- grabbing the graphs to position them.

As was observed in the experiment, participants usually perform pointing in two stages: first they try it when they just reach the starting position, and after a break when participants observe information on the graphs occurs a second stage of pointing. The same two stages can be observed for grabbing. Thus, there are five main sets of activities:

- walking;
- pointing part 1;
- pointing part 2;
- grabbing part 1;
- grabbing part 2.

It is important to mention that three users did not perform grabbing, thus were excluded from the evaluation of grabbing features.

The controllers and the headset movements are tracked through sensors or lighthouses (HTC Base Station in our case), located at a suitable distance. Both controller and headset send some data (like controller inputs) to the computer that processes it depending on the application running. The lighthouses are autonomous and yet need the controller and headset to generate meaningful data, and while the controller and headset do transmit data to the computer, the data would be incomplete if it was not computed with the information coming from the lighthouses [6]. The application used in the experiment was modified by the research group in a way that data that was being transmitted was stored into the CSV files.

During the experiments following data was collected:

- Timestamp;

- Hands:
  - Hand_Right_x, Hand_Right_y, Hand_Right_z (coordinates of the right hand at the moment), Hand_Right_Roll, Hand_Right_Pitch, Hand_Right_Yaw;
  - Hand_Left_x, Hand_Left_y, Hand_Left_z (coordinates of left hand at the moment), Hand_Left_Roll, Hand_Left_Pitch, Hand_Left_Yaw;
- Head:
  - Head_x, Head_y, Head_z (coordinates of the head at the moment), Head_Roll, Head_Pitch, Head_Yaw;
- Eyes:
  - Eye_Right_Openess, Eye_Left_Openess (how open eyes are), Gaze_Orgin_No_Offset_y, Gaze_Orgin_No_Offset_x, Gaze_Orgin_No_Offset_z (gaze offset coordinates);
- Sample (every record for a single participant is marked with a unique Sample label).

5 Data Management

5.1 Data Preprocessing

This section talks about how data was prepared for further analysis and processing after collection.

The differences between the research in question and most other conducted studies are the following: this research has a different motivation, it does not involve two sessions of experiments, typing is not included in the action set, and the experiment for all users is done once. It also does not filter out any runs to avoid learning error. Current research uses five different classifiers and a balanced number of male and female participants.

The application of ML algorithms requires the presence of data in a mathematically feasible format through data preprocessing. DP techniques consist of data reduction, data projection, and missing-data treatment. Data reduction aims to decrease the size of the datasets by means of feature selection. Data projection intends to transform the appearance of the data, like scaling, which scales all features into a predefined same range. Missing-data treatments (MDTs) include deleting missing values and/or replacing them with the estimates [14].

5.1.1 Missing-Data Treatments

Fortunately, the dataset did not include a lot of missing values other than three users who did not perform grabbing activities and two users for whom eye-tracking was not enabled, thus missing-data treatment was narrowed down to excluding these participants from the grabbing and combined task evaluation. After experiments were over, they would ideally result in 37 CSV files containing raw time-series data from sensors, but because of technical issues (users happened to be out of the reach of trackers), for two participants data was split into multiple files. In these cases activity sets were extracted from each file one by one.

5.1.2 Data Projection and Reduction

As part of data projection, having raw time series data files there was a need to split data by the identified features into separate files: walking, pointing part 1, pointing part 2, grabbing part 1, grabbing part 2. To do so, a replay application was created by a research group member: this application loads the raw CSV data and allows to “replay” experience of the participant. Using this application it was possible to note down samples when the user started and finished performing a set of activities for all 37 users.

During preprocessing, the Python pandas library was used. As a result, a dataframe with 37 records identifying edge samples of sets of actions performed was received. With the help of edge samples the data set was cut into five subsets: walking, pointing in two parts, and grabbing in two parts (or three subsets for three participants who did not perform grabbing) and stored into separate files: five or three files per participant. This can be also viewed as data reduction, as arbitrary data was cut off and reduced the size of the dataset.

The next step is to present data in a mathematically feasible format. Using a range loop and a statistics Python library data was reduced from time-series files to a single row of values:

- distance between hands average, minimum, maximum;
- distance between the dominant hand and head average, minimum, maximum;

  distance between the non-dominant hand and head average, minimum, maximum;
- dominant hand pitch, roll, yaw average;
- non-dominant hand pitch, roll, yaw average;
- head pitch, roll, yaw average;
- right eye openness average;
- left eye openness average;

- x, y, and z coordinates of gaze offset.

Distances were calculated as Euclidean distances using coordinates. Steps were repeated for all of the feature sets, thus creating a file for each category: walking, pointing part 1, pointing part 2, grabbing part 1 and grabbing part 2. All five files were saved in CSV format.

Next graphs visualizing data after preprocessing are presented. Figure 5 shows the minimum, average and maximum values of distances between hands during the experiment. Minimum and average distances are pretty similar among the participants, while maximum distance varies highly.

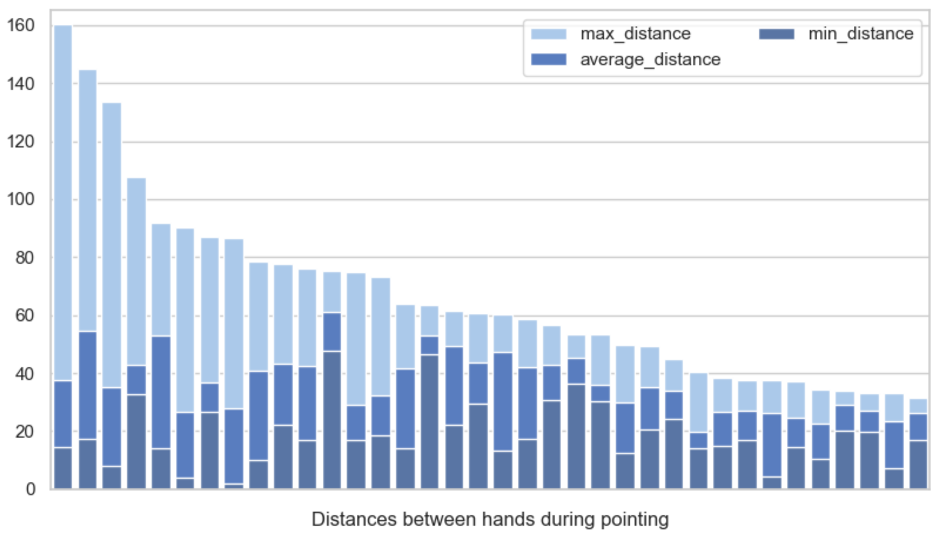

Figure 5. Pointing: Distances Between Hands

The next graph (Figure 6) shows the comparison of average distances between head and hands (dominant and non-dominant) for grabbing and pointing action sets. It highlights that during grabbing hands are closer to the head and are more often located at the same distance from the head. On the other hand, during pointing non-dominant hand is located either much further from the head (being let to hang) or much closer to the head.

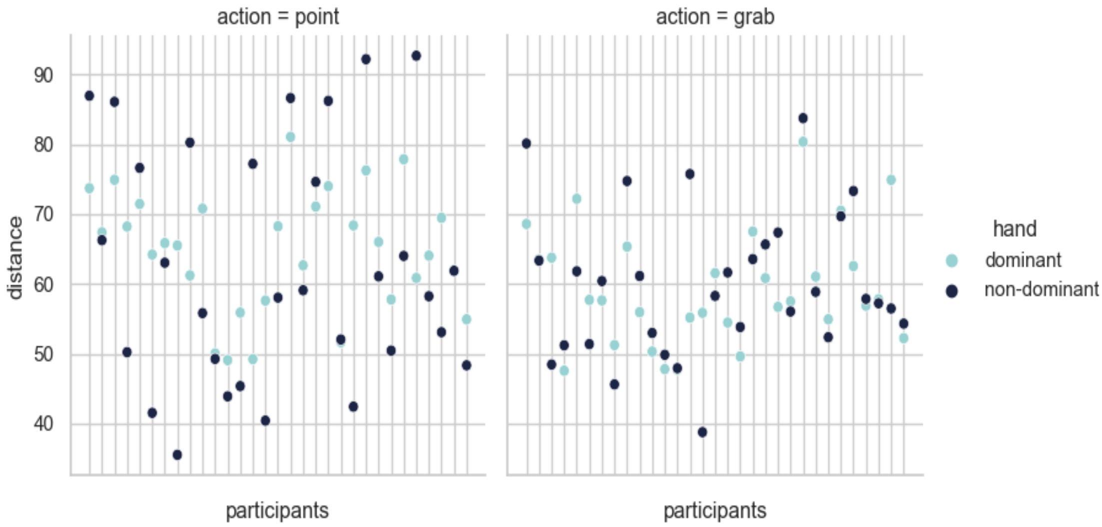

Figure 6. Head to Hands Distance Average

5.2 Data Processing

The following section covers the tools and procedures used during data processing.

In the context of this research machine learning is used to uniquely identify a person from various combinations of subjects. Machine Learning is the study of computer algorithms that improve automatically through experience and using data [11], it is a subset of artificial intelligence. Machine learning plays a key role in research as it enables the extraction of valuable information from a set of data in different forms. There are two types of data analysis categories in Machine Learning: supervised and unsupervised. While using supervised algorithms one must define both inputs and desired outputs. According to the needs, in the data processing stage of this research supervised machine learning algorithms are used. There are also two types of classification problems: classification and regression problems. Regression problems aim to predict numeric values, like quantity, while classification problems are trying to predict categorical values, like labels. This thesis problem is a classification problem, as we are trying to predict labels like “Target user” and “Random user”. Because there are two labels, it is called a binary classification task.

One of the goals of the thesis was to use existing data to try out new classifiers and clustering methodologies, try aggregating the data in a way that has not been applied before. A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more sets of classes. A classifier is an algorithm itself – the rules used by machines to classify data. A classification model, on the other hand, is the result of your classifier's machine learning. The model is trained using the classifier, so that the model, ultimately, classifies your data [14].

Even though walking was included into preprocessing stage, it was not used for identification, as participants walked only once. In this thesis results of the following combinations are compared: pointing, grabbing, and a combined set. Basic steps to get results for these action sets are the same.

To use machine learning algorithms, two datasets are needed: a dataset to train the model, and another dataset, not used previously, to test the model. Pointing stage 1 and grabbing stage 1 feature files are used as train datasets, while pointing stage 2 and grabbing stage 2 are used as test datasets. In the case of the run where pointing and grabbing are combined, datasets are joined correspondingly. This way, after analyzing the results one can tell how accurately users can be identified using the data collected at the beginning of the experience after, for example, a short break. During the initial tests when choosing optimal parameters for classifiers they were trained and tested on the groups of users of sizes 2, 5, 10, 15, 20, 25, 30, and 35, resulting in 122 tests per time. In the case of grabbing the last group's size is 34, and after eye-tracking features were introduced size of the dataset went down to 32 users for a combination of grabbing and pointing because eye tracking was not working for two users.

Once train and test datasets are loaded processing starts: code runs for all the group sizes and picks up a corresponding amount of users. After, the classifier is trained for each user by using that user's feature points as positive examples and all the feature points from other users as negative examples[11]. For example, for a group size of two, we will have two runs: in the first run one user is marked as target user, in the second run the second user is marked as a target user. This is the one-versus-all strategy used for multi-class classification [11]. Once the classifier is trained, the next step is to use it to identify a previously unobserved test dataset.

In the research, we use five classifiers: Logistic Regression classifier, Support Vector Machine classifier, Random Forest classifier, KNN classifier, and Decision Tree classifier. As can be observed in the following section, the KNN classifier shows the best results. Following sections give short overview of used classifiers.

5.2.1 Logistic Regression Classifier

Logistic regression classifier belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. The advantage of this classifier is that it is fast and relatively uncomplicated. Although it is used primarily for binary classification, like in our case, it can also be applied to multiclass problems [17].

One of the important parameters for sklearn Logistic Regression linear model is the solver. A solver is an algorithm to use in the optimization problem. In this case, ‘liblinear’ ‘newton-cg’, and ‘lbfgs’ solvers were tested. ‘Liblinear’ is used for small datasets, like the current one, ‘lbfgs’ is used when more efficient computation is needed. According to the test dataset, the most efficient solver in the case of this particular problem is ‘liblinear’ (Figure 7).

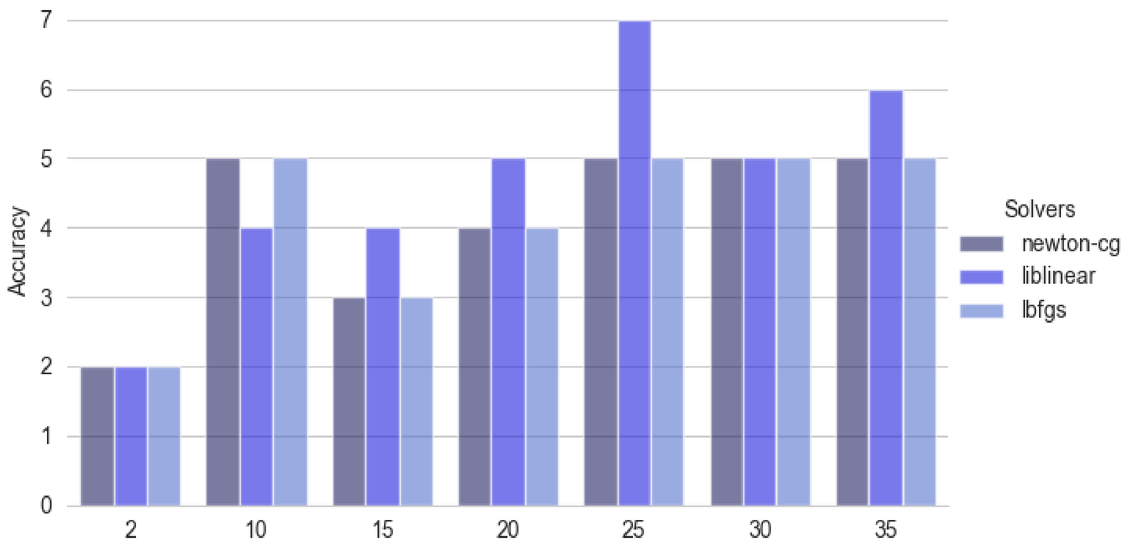

Figure 7. Logistic Regression Solvers

5.2.2 Support Vector Machine Classifier

Support Vector Machine (SVM) classifier has the following advantages: it is effective in high dimensional spaces as it is memory-efficient because it uses a subset of training points in the decision function (called support vectors) [18]. Kernels are mathematical functions used in SVM. Linear kernel works best when there are a lot of features, as in the case of our dataset, the polynomial kernel is a generalized representation of the linear kernel and is used rarely. Gaussian Radial Basis Function (RBF) is used for non-linear data. The sigmoid kernel is used in neural networks [19]. As this dataset has lots of features linear kernel is the preferred option (Figure 8).

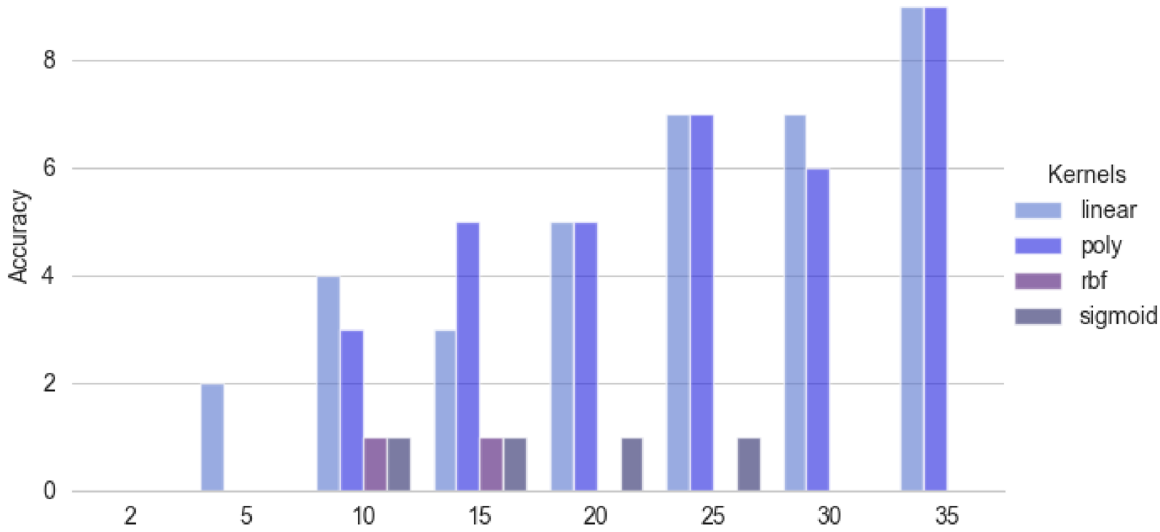

Figure 8. SVM Kernels

5.2.3 Random Forest Classifier

Random forest is one of the most used algorithms, because of its simplicity and diversity: it can be used for both classification and regression tasks. The algorithm builds a “forest” of decision trees based on the input data [20]. The number of trees in the forest can be controlled through the n_estimators parameter: usually the more trees – the better results, but it is recommended to stop when result does not meliorate with adding more estimators. Keeping a number of estimators below 10 is the right choice in terms of the current dataset (Figure 9). Five and ten estimators performed better than any higher number. Additionaly, keeping number of estimator low helps to improve the performance.

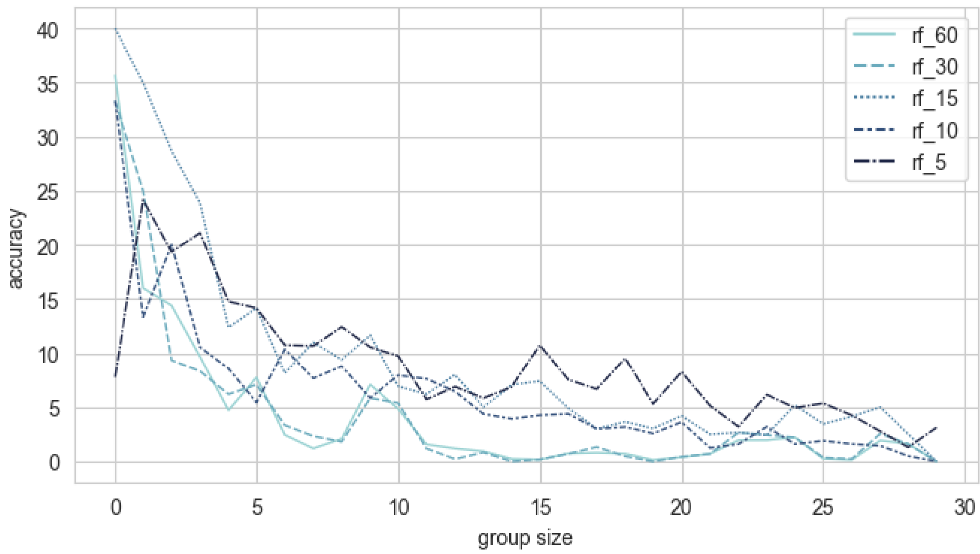

Figure 9. Random Forest Estimators

5.2.4 Decision Tree Classifier

Decision Tree is a frequently used machine learning algorithm. A tree structure is constructed so that it breaks the dataset down into smaller subsets eventually resulting in a prediction [21]. Testing parameters did not result in any significant changes in the accuracy.

5.2.5 K-Nearest Neighbours Classifier

KNN is one of the simplest Machine Learning algorithms, yet in our case, it showed the best results. K-NN algorithm is called a lazy learner algorithm because at the training phase it does not learn – it just stores the dataset and when it gets new data, then it classifies that data into a category that is most similar to it.

KNN algorithm steps are the following:

- Select the number K of the neighbours;
- Calculate the distance of K number of neighbours;
- Take the K nearest neighbours as per the calculated distance;
- Among these k neighbours, count the number of the data points in each category;

- Assign the new data points to that category for which the number of neighbours is maximum.

There is no way to determine the best value for "K", so it is needed to try some values to find the best out of them [22]. Table 1 presents the results of testing values 1 and 2. In all the following manipulations number of neighbours is 1, as it shows significantly better results.

Table 1. K-Nearest Neighbours

No	2	5	10	15	20	25	30	35
1	2	1	6	5	9	12	10	11
2	0	0	0	0	0	0	0	0

Another important parameter is a distance metric. The distance metric is responsible for setting the distance metric to be sued for the tree. In the research, the Minkowski metric is used. When parameter p equals 1, this is equivalent to using Manhattan distance, when p is 2 – to Euclidian. Euclidian distance shows better results (Figure 10).

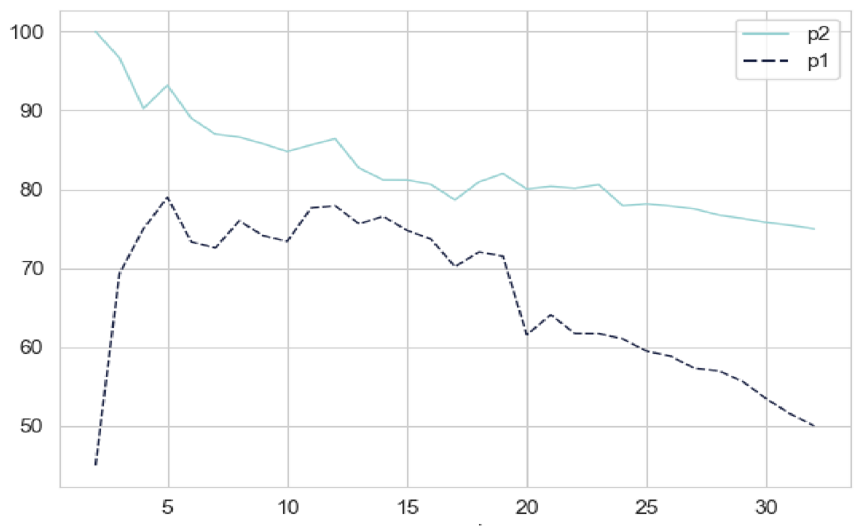

Figure 10. KNN Distance Metric

5.3 Metrics

In the final results of this research classification accuracy is reported – the number of correct classification decisions divided by the total number of decision (i.e. number of test cases). Classification performance across group sizes is evaluated as follows: for each group size (from 2 to 32) all possible combinations are found, then a sample of size 100 is drawn from them. For each such draw, classifier is then evaluated on this subset of participants [5]. For example, for a group size of 2, there are 200 runs: in the first run user one is marked as target user, in the second run the second user is marked as a target user, and this is repeated 100 times for every sample of random 2 users picked from the dataset. Accordingly, 300 runs for a group size of 3 users, 400 for 4, and so on. For a group size of 31, there are only 32 possible subsets and for group size 32 there is only one, full set.

During each run there is a training dataset and one true array of labels: out of N number of participants one, target participant, is marked by label "1" and the rest, random participants, are marked by label "0". Using train data, classifier tries to create a similar array of labels but already for the unknown test dataset. Classifiers can give one of the following results: false-positive, false-negative, true-negative, and true positive. False-positive means that the classifier labelled random participant as target participants, false negative means that target participant was not recognized and marked as random, true positive result signifies that target user was identified correctly and true negative – that random user was identified as random. A run is marked as successful if only true results are received, either positive or negative. For example, if the group size is 30, successful result would mean 29 true-negatives and 1 true-positive. These values can be represented through the classification matrix. An example of the classification matrix from the KNN classifier is presented below (Figure 11), output shows 0 false-negative results (lower left corner) and 0 false-positive results (upper right corner).

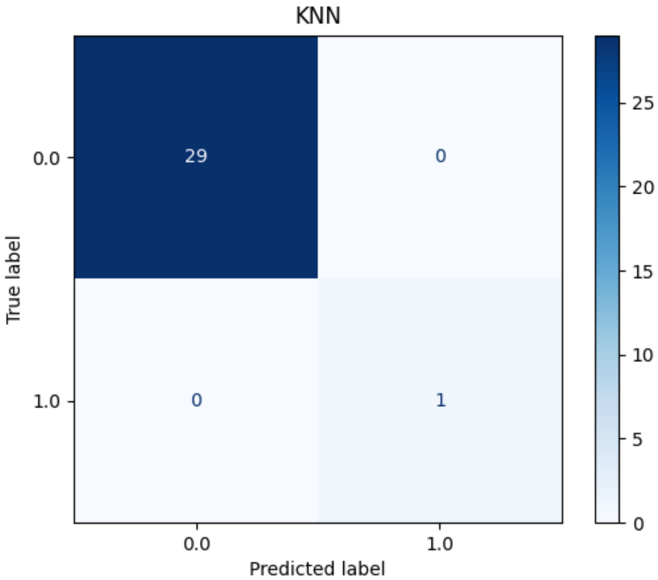

Figure 11. Example Classification Matrix

6 Results

6.1 Experimental Results

First, looking into results (Figure 12) we can observe that combination of grabbing and pointing reports better results than separate features: a combination of features is 10% more accurate than pointing. Pointing results are on average 10% better than grabbing results because pointing tasks allow users to exhibit more freedom in the manual movement [5].

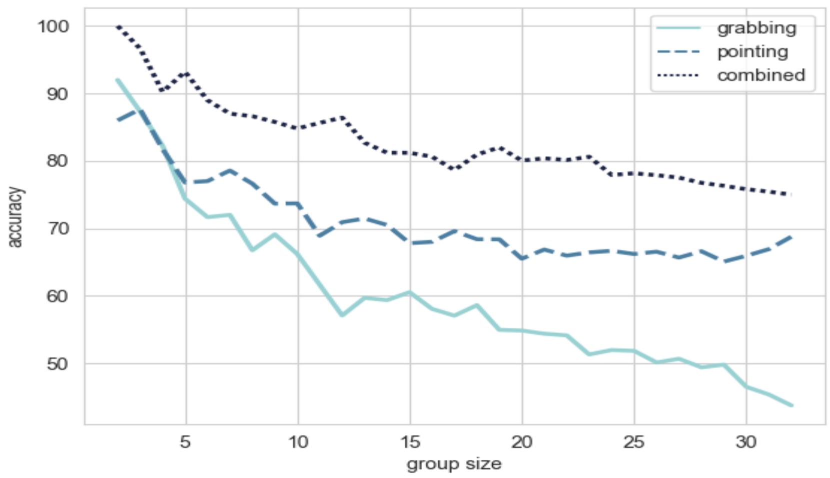

Figure 12. Comparison of Action Sets

When it comes to the comparison of classifiers, the K-Nearest Neighbours classifier is the top leading one in this research. A possible reason for this algorithm to perform so well is due to the fact that it is a “lazy learner”, thus the model is not biased against the target user when the record comes. The next best classifier, underperforming KNN by almost 20%, comes Support Vector Machine, and then comes logistic regression with about 7% difference. Even though Support Vector Machine and Logistic classifiers are explicitly used for binary classification problems, they did not show the best results for the current problem. This might be caused by the fact that usually the class for the normal state is assigned the class label 0 and the class with the abnormal state is assigned the class label 1, like a target user, so in case of this problem, we do not have enough examples of abnormal state class members – it always equals to one. Logistic Regression is most useful when you want to understand how several independent variables affect a single outcome variable, which is not the case: most of the variables in the dataset are independent [23]. Random Forest and Decision Tree proved to be least efficient for this classification problem.

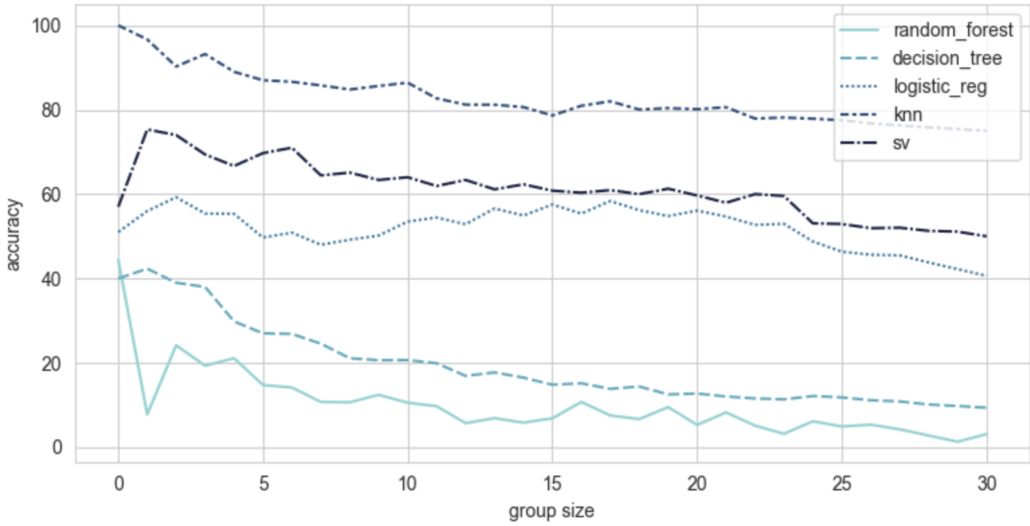

Figure 13. Comparison of Classifiers

The following graph (Figure 14) shows the difference between classification accuracy using the same dataset, but a slightly different set of features: selected feature graph shows results when feature set consisted of average values, excluding any minimum and maximum values. The assumption is that people share the edge cases of movement: for example the likelihood of reaching hands out in virtual environments is high.

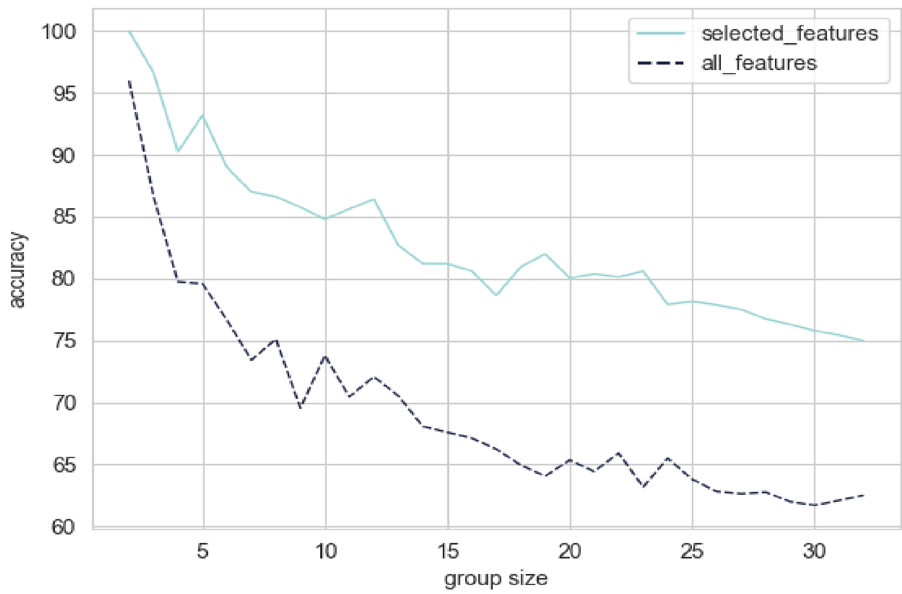

Figure 14. Feature Selection

Results of identification can be enhanced by further researching the combinations of classifiers, like neural networks, features, like finding standard deviations instead of using maximum and minimum values. Among others, accent on eye-tracking can enhance the results.

6.2 Impact

Both Availability and Confidentiality are the crucial pillars of Cyber Security. Clearly, Virtual and Augmented Reality raises the question of priorities: on one hand, data is an essential part of the Virtual reality experience. In future applications of this technology can go as far as the use in medicine, military or any other field where people's lives rely on the precision and uptime of applications. Without a doubt, availability has to be ensured. On the other side, user's right for Confidentiality must not be neglected. As it has been discussed in the Section 2, biometric aspect of data collected in VR differentiates it from the data collected from other devices in the daily use, and thus it has to be treated as highly sensitive. In this case data is being at the same time the solution to the problem of ensuring Availability and an obstacle to ensuring Confidentiality.

While it is not feasible to avoid collection of data in Virtual Reality, it is possible to communicate to users in a transparent manner what and why is being collected, how data is stored, who has access to it, and, among others, how sensitive data is. As an example, Facebook offers virtual, mixed, and augmented reality ("XR") hardware and software products and operates a platform under the Oculus brand [28]. As stated on the official website [28], Oculus collects information about user’s physical movements and dimensions when they use an XR device. For example, if hand tracking feature is enabled, they collect technical information like estimated hand size and hand movement data to enable this feature. This research proves that users performing trivial tasks can be identified uniquely using such data. Oculus Privacy policy makes it quite clear how the information is used, and who it is shared with, and, owing to GDPR, there is also a clear procedure for deletion and retention of data. What stays unclear to a regular user in this overall common privacy policy notice is how sensitive the information collected is. Right or wrong, in present time such knowledge user has to obtain conducting independent research, and not from the policy makers.
Confidentiality refers to our ability to protect our data from those not authorized to view it [29]. Obviously, it is impossible to protect data in case user is not aware of what such data ultimately represents. As a result, people unknowingly share information they might have intended to keep confidential.

7 Summary

Virtual and Augmented reality is expanding on a daily basis, becoming more important and dominant in the everyday life of an increasing amount of population. Data collection during the experience in Virtual Reality is essential for the quality of experience, as VR/AR technologies are highly dependent on it.

This research established that data collected during a trivial experiment in a Virtual Reality environment performing common actions can be used to later identify a participant among a group. Data was collected from 37 participants during an educational experiment with a 3D crane. The results from five classifiers with various parameters were compared, finding the best fit for this particular problem. Models are trained on two action sets: pointing and grabbing. Results presented in this research show that the K-Nearest Neighbors classifier achieves the best identification rate, while the best feature set for identification is the combination of pointing and grabbing.

The accuracy rate of identifying a human being of solid 75% was achieved for the group of 32 people. As the experiment was kept extremely close to the real-life experience, such results mean that data collected in Virtual Reality is both personal and sensitive according to GDPR. Even though manufacturers and lawmakers are notifying users about what kind of data they are collecting, private information can be inferred from it, threatening privacy, anonymity, and autonomy of participants and being an obstacle to ensuring their Confidentiality. The author believes, that individuals consenting to their data collection and processing need to be aware of what kind of risks such data collection introduces, as it is impossible to protect our information if it stays unclear what it ultimately represents.

References

[1] McKinsey&Company. (2017). “Augmented and virtual reality: The promise and peril of immersive technologies,” [Online]. Available : https://www.mckinsey.com/industries/media-and-entertainment/our-insights/augmented-and-virtual-reality-the-promise-and-peril-of-immersive-technologies

[2] MarketsMedia (2017). “AR Becomes a Reality in Trading,” [Online]. Available : https://marketsmedia.com/ar-reality-trading/

[3] Grand View Research. (2021). “Virtual Reality Market Size, Share & Trends Analysis Report By Technology (Semi & Fully Immersive, Non-immersive), By Device (HMD, GTD), By Component (Hardware, Software), By Application, And Segment Forecasts, 2021 – 2028,” [Online]. Available : https://www.grandviewresearch.com/industry-analysis/virtual-reality-vr-market

[4] J. Dalton and J. Gillham, “Seeing is believing, how virtual reality and augmented reality are transforming business and the economy,” PWC, 2019

[5] K.Pfeuffer, M. Geiger, S.Prange, L.Mecke,“Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality,” in CHI Conference, 2019

[6] Alessandro Mirani, “User Behavior Analysis for Predictive Virtual Reality Applications: An Ethical and Data Security Perspective,” [Online]. Available : https://a-lab.ee/edu/theses/defended/2110

[7] A. Kupin, B. Moeller, Y. Jiang, N. K. Banerjee, and S. Banerjee, “Task-driven biometric authentication of users in virtual reality (VR) environments,” in International conference on multimedia modeling, 2018

[8] GDPR.EU. “What is considered personal data under the EU GDPR?” [Online]. Available : https://gdpr.eu/eu-gdpr-personal-data/

[9] CNBC. “Coronavirus could be catalyst to reinvigorate virtual reality headsets”. [Online]. Available : https://www.cnbc.com/2020/05/02/coronavirus-could-be-catalyst-to-reinvigorate-virtual-reality-headsets.html

[10] Statista. “Augmented reality (AR) and virtual reality (VR) headset shipments worldwide from 2020 to 2025”. [Online]. Available : https://www.statista.com/statistics/653390/worldwide-virtual-and-augmented-reality-headset-shipments/

[11] T. Mitchell, “Machine Learning,” New York, USA, 1997, [Online]. Available : http://www.cs.cmu.edu/~tom/mlbook.html

[12] C.E. Rogers, A.W. Witt, A.D. Solomon, K.K. Venkatasubramanian, “An approach for user identification for head-mounted displays,” In: ACM ISWC, 2015

[13] MonkeyLearn. “Machine Learning Classifiers - The Algorithms & How They Work”. [Online]. Available: https://monkeylearn.com/blog/what-is-a-classifier/

[14] J. Huang, Y. Li, M. Xie. “An empirical analysis of data preprocessing for machine learning-based software cost estimation”. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584915001275

[15] K. Bye, D. Hosfelt, S. Chase, M. Miesnieks, T. Beck, “The ethical and privacy implications of mixed reality”. [Online]. Available: https://dl.acm.org/doi/10.1145/3306212.3328138

[16] E. Dick, “Balancing User Privacy and Innovation in Augmented and Virtual Reality”. https://dl.acm.org/doi/abs/10.1145/3306212.3328138

[17] Real Python: “Logistic Regression in Python”. [Online]. Available: https://realpython.com/logistic-regression-python/

[18] ScikitLearn, “Support Vector Machines”. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html#

[19] Dataspirant, “Seven most popular svm kernels”. [Online]. Available: https://dataaspirant.com/svm-kernels/

[20] Builtin: “A complete guide to the random forest algorithm” . [Online]. Available: https://builtin.com/data-science/random-forest-algorithm

[21] B. A. Keen, “Decision Tree Classifier in Python using Scikit-learn”. [Online]. Available: https://benalexkeen.com/decision-tree-classifier-in-python-using-scikit-learn/

[22] Javatpoint, “K-Nearest Neighbor(KNN) Algorithm for Machine Learning” . [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[23] Machine Learning Mastery, “4 Types of Classification Tasks in Machine Learning”. [Online]. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/

[24] W. Barfield, M. Blitz, “Research Handbook on the Law of Virtual and Augmented Reality”.

[25] Sheri B. Pan “Get to know me: protecting privacy and autonomy under big data's penetrating gaze”, Harvard Journal of Law & Technology, Volume 30, Number 1, 2016

[26] JdSupra, “Virtual reality: top data protection issues to consider”. [Online]. Available: https://www.jdsupra.com/legalnews/virtual-reality-top-data-protection-43864/

[27] A. Tepljakov, EuroVR 2019 Conference Keynote. [Online]. Available: https://recreation.ee/at/eurovr2019/#/13

[28] “Oculus Privacy Policy”. [Online]. Available: https://www.oculus.com/legal/privacy-policy-for-oculus-account-users/

[29] J. Andress, “The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice”. [Online]. Available: https://books.google.ee/books?id=9NI0AwAAQBAJ&pg

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Oleksandra Zamana

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Personal identification in Virtual and Augmented Reality: Control Systems Perspective", supervised by Aleksei Tepljakov and Saleh Al-saleh.

   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Final Results

Table 2. Final Results

group	random_forest	decision_tree	logistic_reg	knn	sv
2	44.5	40	51	100	57
3	7.777778	42.33333	56	96.66667	75.33333
4	24.16667	39	59.25	90.25	74
5	19.33333	38	55.4	93.2	69.4
6	21.11111	29.83333	55.33333	89	66.66667
7	14.7619	27	49.71429	87	69.71429
8	14.16667	26.875	50.875	86.625	71
9	10.74074	24.55556	48	85.77778	64.44444
10	10.66667	21.1	49.2	84.8	65.1
11	12.42424	20.63636	50.18182	85.63636	63.36364
12	10.55556	20.66667	53.5	86.41667	64
13	9.74359	19.92308	54.46154	82.69231	61.92308
14	5.714286	16.92857	52.85714	81.21429	63.35714
15	6.888889	17.73333	56.6	81.2	61.13333
16	5.833333	16.5	54.9375	80.625	62.3125
17	6.862745	14.82353	57.52941	78.64706	60.82353
18	10.74074	15.16667	55.38889	80.94444	60.33333
19	7.54386	13.84211	58.36842	82	60.94737
20	6.666667	14.4	56.2	80.05	60
21	9.52381	12.52381	54.80952	80.38095	61.28571
22	5.30303	12.72727	56.09091	80.13636	59.68182
23	8.26087	12.04348	54.69565	80.6087	57.95652
24	5.138889	11.58333	52.70833	77.91667	60
25	3.2	11.36	52.96	78.16	59.56
26	6.153846	12.15385	48.76923	77.88462	53.07692
27	4.938272	11.81481	46.37037	77.51852	52.92593
28	5.357143	11.14286	45.64286	76.75	51.92857
29	4.252874	10.86207	45.48276	76.31034	52.03448
30	2.777778	10.13333	43.8	75.8	51.3
31	1.310484	9.778226	42.2379	75.44225	51.10887
32	3.125	9.375	40.625	75	50