Investigating a gesture based interaction model, controlling a truck with the help of gestures

Niclas Brynolfsson

Department of Applied Physics and Electronics
Umeå University, Sweden

Abstract. This paper examines if a gesture based interaction model to control a heavy vehicle is a suitable mean of control. It looks at what the interaction model needs to make it reliable, trusted and usable. The methods that was used during this thesis included a literature review, prototype development and testing. To test this a prototype was developed in Unity. To handle the gesture recognition, it was paired with a Kinect 2.0, and the gesture learning was done with Visual Gesture Builder. With this we created two sets of gestures that were tested against each other.

We found that a the gesture based interaction model was found intuitive, natural and seems to allow the operators to efficiently execute their tasks. However, there are two key aspects that needs to be taken in to consideration. Firstly the gesture recognition has to be at a really high standard as in having high recognition and well thought out gestures, both out of safety reason, but also to not cause frustration with the users. Secondly, the gestures themselves need to be designed in a way that will not cause fatigue, with poorly designed gestures the users would not be able to use this interaction model for a longer period of time.

Abstract. Det här examensarbetet undersöker ifall en gestbaserad interaktionsmodell för att styra tunga fordon är ett lämpligt sätt. Det undersöker vad interaktionsmodellen behöver för att vara säker, pålitlig och användbar. Metoderna som användes i examensarbetet inkluderar en litteraturgranskning, prototyputveckling och användartester. För att kunna testa frågeställningen så utvecklades en prototyp i Unity. En Kinectkamera 2.0 användes för att spela in och tolka gesterna och gestinlärningen gjordes med hjälp utav Visual Gesture Builder. Med detta skapades två set med gester, en på fyra gester och en med sex gester.

Resultaet av arbetet var att en gestbaserad interaktionsmodell ansågs vara intuitiv, naturlig och tillät användarna att använda prototypen på ett bra sätt. Däremot är det två huvudaspekter som måste tas i beaktning. Det första är att gestigenkänningen behöver hålla en extremt hög standard, så som att den ska ha en hög igenkänning och väl genomtänkta gester, både ur en säkerhetsaspekt och för att förhindra frustration hos användaren. För det andra, gesterna i sig behöver vara designade på det sättet att de inte orsakar trötthet och krämpor i musklerna. Med dåligt designade gester kan inte användaren nyttja interaktionsmodellen någon längre period utan att det blir för påfrestande.

Table of Contents

1 Introduction

Human gesturing has been a means of communication since the beginning of mankind. Some researcher claims that the spoken language originates from gestures [1]. We use it for everyday application, wave hello to friends or point in a direction. It can also be used for helping someone backing up with a trailer or when directing airplanes at the airport. All in all, an universal language that could be understood by anyone. But how would this translate into Human-Machine-Interface (HMI)? The automotive industry has evolved and autonomous vehicles becomes more and more prominent. As it is now there is little knowledge about the interaction design for self-driving vehicles, especially for users that are controlling when standing outside of the moving vehicle. What if the autonomous system reaches its limit and cannot continue to drive on its own, what should the human interaction with for instance a truck be like? As users often are habit driven [2], it could be hard for developers to implement new features or completely new ways of interacting, that the end users accepts and are willingly to use. What would it take to break those habits, and implement new features?

In the hub [3] is a project by Scania, in collaboration with RISE, Boliden and ICE makers. The project aims to explore how people should interact with driverless vehicles in a connected transport system. The goal of the project is to design new interaction models that will be used for efficient and safe transports, "In The Hub". There are three use cases for the project. This master thesis will look at one specific, user case A. User case A is the natural interaction between an automated vehicle and the user when standing next to it. What would the interaction model look like when you need to solve a problem fast and optimize the flow of traffic? The work is focused to develop a limited amount of commands that works on a global level and independent of user context. Could an interaction model based on gestures interpreted by a computer be a suitable model?

Research done on gesture control can be dated all the way back to the 80's [4]. Since then the technology has improved tremendously. Gesture sensing hardware as become widely available, through products such as Microsoft Kinect [5], Leap Motion Controller [6], Nintendo Wii [7] or Google Soli [8]. In addition, if combined with machine learning, especially if you use a finished solution like Visual Gesture Builder [9] that is included with the Kinect developer kit, the development time is drastically reduced to implement a gesture database. With these improvements, how would an interaction between an autonomous vehicle and a user look like? This is what the thesis aims to find out. It will investigate how a gesture based interaction model could look like to control an autonomous vehicle when standing outside, next to it.

This thesis will consist of four main parts. A background and theory, where important background information will be presented to get a greater understanding for this thesis. A method that will present what we did and how we did it to develop the prototype and test it. The result given from the testing and a discussion were we will go through the result, and give our own opinion on the matter.

1.1 Objective

The aim of this thesis is to research, implement and test interaction models to control a truck when standing outside, next to it.

The objective is to validate if a gesture based interaction model is a viable method to control a heavy vehicle, and look at what would be needed in this gesture based model to make it reliable, trusted and usable. To achieve this goal, a prototype will be made with two sets of gestures which will be tested against each other, to see which one is more suitable, and gain knowledge about what to look for. This can later be used as guidelines for future development of gesture based interaction.

2 Background & Theory

In this section some background information will be given. We will present different hardware for tracking gestures, software that is used for testing and some literature review to gain a greater understanding on the subject. It will also cover the tools used in this project.

2.1 Rise

This thesis has been carried out at Rise, which is a Swedish state-owned research institute, collaborating together with universities, industry and the public sector to drive international competitive sustainable development and growth. In January 2020 Rise had around 2800 employees over five organisational divisions. Piteå consists of two divisions: the User Experience Unit at RISE Research Institutes of Sweden and RISE Energy Technology Center.

2.2 Hardware & Software

To interpret gestures there are some hardware and software that would be suitable and helpful. This subsection will present some of them.

2.2.1 Kinect Kinect is an advanced sensor with depth and infrared camera and voice recognition. The first Kinect-device was launched by Microsoft on November fourth 2010 [5] and was a huge success for Xbox 360 and gaming. The In 2013 the success was followed with the Kinect 2.0 [10] However, this version did not become as big of success as the first version, and on 25th of October 2017 Microsoft announced that they would discontinue the production of the camera [11]. Nevertheless, even though the second generation of the Kinect was not as huge of success as the first generation, the use case for it would far exceed gaming. In 2014 the official adapter for Windows was released, and with it the official Software Development Kit (SDK). At a price point of around US$100 it brought depth camera to the masses and the uses for it in academic works was popular. Its specifications [12] consists of a resolution of 1920x1080 for the color camera, and 512x424 for its depth camera. The framerate is 30fps and it has a max range of 4.5m. The field of view for depth is 70 degrees horizontal and 60 degrees vertical. There is an official SDK for the Kinect available, but it is no longer actively updated. On 24th of February, 2019, Microsoft unveiled their next version of Kinect, Azure Kinect DK. Instead of gaming as main focus, this device is solely aimed for the professional use case. It has a smaller form factor, improved camera and can utilize Microsofts Azure cloud service.

2.2.2 Leap Motion Leap Motion is a real time tracking device for hand gestures. It has a field of view of 150 degrees to the side and 120 degrees in depth with a range of roughly 80 centimeters and a resolution of up to 200 frames per second [14]. It was launched on June 22nd 2013 [6]. The software

Fig. 1. Microsoft Kinect 2.0, launched in 2014. Image taken from *Polygon* [13].

development kit can be found on their website and it supports all major operation systems. There have been some testing done with gesture recognition and sign language, specifically Australian sign language [15]. The paper showed that the Leap Motion is capable of accurately recognizing basic signs. However, when it comes to more complex sign, or the ones that require a broader field of view, for instance when the hands are close to the body or head, the Leap Motion struggles. This is a limitation of the field of view, but also that it can not see the hands when something is obstructing them. For instance if one hand is above the other. Still with these limitations in mind the Leap Motion could be a powerful tool when used for recognition because of its high resolution.

Fig. 2. Leap Motion USB Device with its package. It is compatible with all major personal computer operation systems and use IR-cameras to capture hand movement. Image taken from Leap Motions website [16].

2.2.3 Visual Gesture Builder Visual Gesture Builder (VGB) [9] is a program that is included in the Kinect developer kit. With this you are able to map out a gesture, using either discrete (static) or continuous (movement) and have the program use machine learning train to recognize that gesture. Since this program does not require the user to write a single line of code, the development time is reduced. To use VGB, the user first need to record some video clip. This is done with the help of Kinect Studio, also included in the Kinect developer kit. These clips are then loaded into VGB were the clips are tagged, as in you mark when the gesture is performed, see figure 3 for an image of the application. The blue lines at the bottom of figure 3 represent when a gesture is executed, tagged as true. Everything else is tagged as false. The discrete gesture use AdaBoost-Trigger [17] as an algorithm for learning the gestures. The AdaBoost-algorithm works in the way that it combines several weak classifiers to make a strong classifier. A weak classifier performs poorly, but better than guessing at random. Therefore when adding several weak classifiers you get one strong. The advantage of AdaBoost is that it is fast, simple and is less susceptible to overfitting. To me more precise, to give to much data and therefore the outcome will be bad and the confidence of recognizing gestures will be low. Disadvantage of AdaBoost is that it is vulnerable to noisy data. If one of the recorded gestures was badly recorded, as in that the skeleton tracking is corrupted or not correct, the outcome will be poor and the confidence level of recognizing a gesture will be low.

2.2.4 Unity Unity is a 2D/3D engine that allows for real-time development [18]. This allows for fast iterations since the developer for instance can run a game or other projects they are developing inside of the Unity-tool and there are no need of lengthy compiling time after a change has been made. Furthermore, this visual editor allows the creators to do simple drag and drop programming, and with support for C#-script [19] there are huge possibilities for customization. In addition, the Unity assets store further allows for faster development speed since there are a lot of pre-modeled environments and props. However, Unity has no built in 3D-modeling tool, for this other software tools are required. In addition with the pre-models, the assets store can also offer finished solution to connect different hardware with Unity, such as VR-headsets, Kinect, Leap Motion Controller or other game controller. These can then be modified so they suit the application. Unity is free to use and publish projects, at least until you make sells exceeding US$100 000. After that a subscription to Unity is needed.

2.2.5 MediaPipe Holistic MediaPipe Holistic [20] is real-time tracking software for human pose, face landmarks(key points on the face, such as nose, mouth and etc.) and hand tracking on mobile devices and desktop. There is no need for expensive depth-cameras, the only thing needed is a camera phone or web camera and skeleton tracking, face tracking and gesture recognition is available. This opens up the ability to do gesture recognition on the cheap. Combine this with machine learning and complex gesture detection could be achieved.

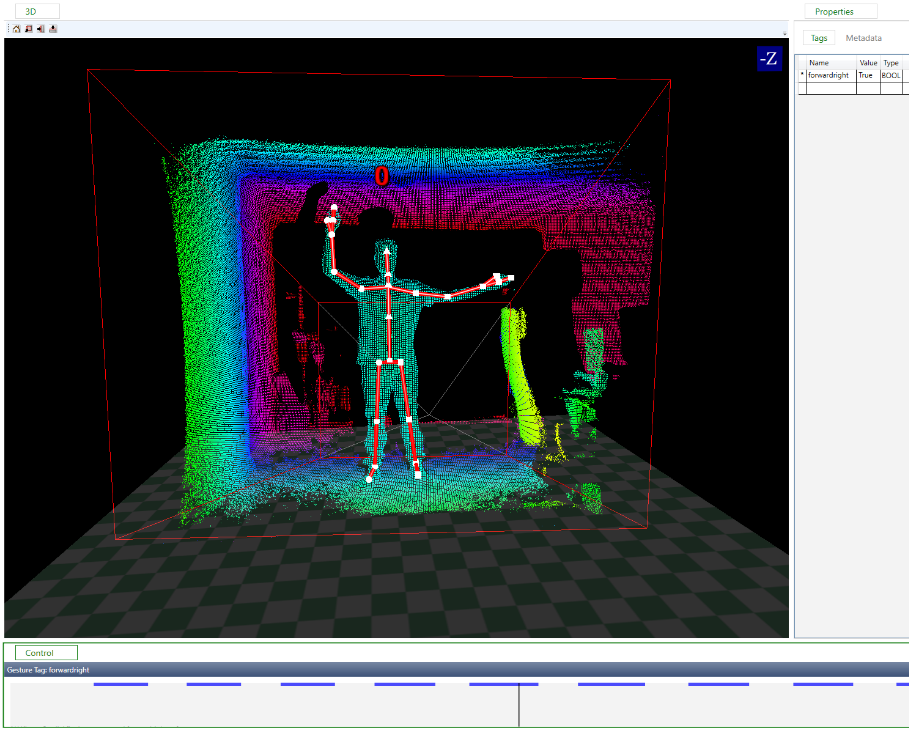

Fig. 3. Visual Gesture Builder, skeleton tracking of the human and the blue bar representing a tagged gesture.

2.3 Machine Learning

Machine learning has been around for a long time. The first documented case of machine learning was in 1959, "Some Studies in Machine Learning Using the Game of Checkers" [21]. Arthur L. Samuel wrote a chess program that would learn and improve after each game it played, either against another computer or real player. In short, it is a method for the computer using experiences to improve its performance. There are a number of ways that machine learning is implemented and trained [22].The three main ones is supervised learning, unsupervised learning, semi-supervised learning. From these there are other variants but almost all of the is based on these three. Supervised learning is when you give the learner a set of labeled and tagged data as training. This is the most common way to implement machine learning. Unsupervised learning is when the learner is only given unlabeled data and makes its predictions for all unseen points. For example,this can be used to find patterns in unknown data or clustering, split the dataset into groups based on their similarities. Semi-supervised learning is when the learner receives both labeled and unlabeled data, and then makes prediction on the unlabeled data.

2.3.1 Speech Recognition Speech recognition is not something new and has been used for many years. However, in recent year with the addition of Google Home and Amazon Echo it has become far more common and would not be uncommon to find in many homes. But with speech recognition comes other problems, one of them is reliability. Since every person is different, they talk different, may it be dialects or different languages. This is a problem, specially in an industrial environment where many operations are of a critical meaning. Room for error is big.

Komiya, Kayoko et al. conducted a research [23] in 2000 where they tried to control a wheelchair with the help of voice to determine if it is an viable method of control, when the user could not use their hand to control a joystick. They made a small robot to mimic a wheelchair, where the participants could control this either via a keyboard, or with their voice. The robot also had two ways of going forward. In latched mode it continued to move forward until the next command and the second was momentary, an on/off mode. The course that the participants had to navigate around was a figure eight. The result from this test was that when controlling with voice, a multiple amount of commands had to be given, in both modes, resulting in heavy load on the user and slow speeds compared to the keyboard. However, the result was promising for controlling with voice but require more development.

RISE Piteå have done an unpublished pilot study on future use of voice commands in a industry. They built a small prototype where a user could control a Lego model consisting of cart and a gate, via voice. They conducted user tests both on site and online. The test showed promising result, where users could see potential in voice commands. Everyone participant in the study were positive and could see themselves using voice interaction for communicating with machines and computer system in an industrial environment.

2.4 Gestures

This subsection will go through some previous research and also how gestures can be used today, this is to get a better understanding for gesture based interfaces.

2.4.1 Air Marshalling Air Marshalling [24] is the act of directing airplanes at an airport, aircraft carrier or helipad. This is done with the help of a standardized set of gestures that will direct the aircraft or helicopter, see figure 4. These are developed to be universal and easy to understand so that it does not matter at what airport the pilot are, they would still understand them. This is necessary because flights are coming from all around the world, and therefore there is a need for a universal language that is not dependent on spoken directions.

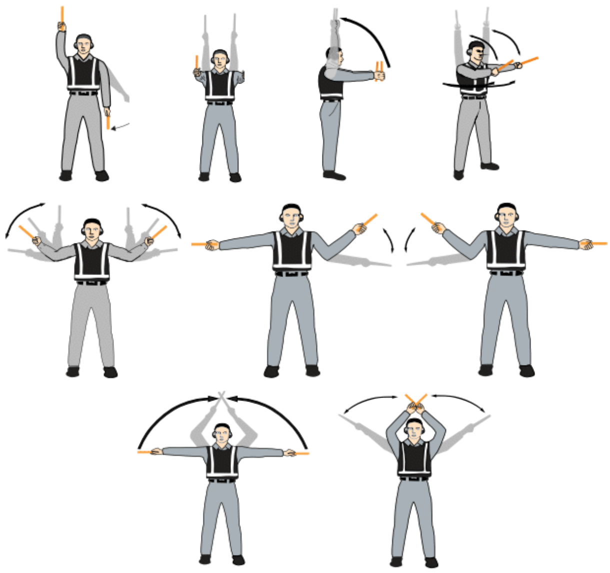

Fig. 4. Image taken from *Fear of Landing* [25] showing gestures used for air marshalling.

2.4.2 Gesture Based Interface To control a interface with the help of gestures, gestures can be divided into two categories, small and big. Small would be gestures done with hands or fingers, gestures that could be executed sitting down or in a smaller confinement. Larger gestures would be something that requires the whole upper or complete body, but also a larger environment to perform these gestures.

David Procházka et al. [26] wrote that there are several aspect when developers design their control mechanism based on gestures. One key aspect is simplicity, the gesture can be neither long or complicated. Another point is that it should be natural for the user. For instance like turning a book page. They also said that, if possible, the gestures should not be designed by the developer, but the target user group. The third key aspect was the implementation. Good performance is important. If there is no immediate feedback, whether the gesture was correctly done, the user will be confused.

The military of United States of America ordered a study on gesture-based controls for robots. Linda R Elliott et al. [27] investigated if a gesture based solution, either via camera or wearable systems could be a potential use case in the future of US military. They explored how ground robots could be controlled by soldiers. Some of the issues they ran into when dealing with a camera-based system is the problem with field of view, line of sight and obstruction of the camera, such as fog, smoke or bad weather. However, that could be mitigated with the help of infrared or thermal cameras. But in conclusion, as of year 2016 the report deemed that a camera-based system had too many short comings to be deployed as a useful system, but with improvements in the future, it could become an alternative to controlling robots when in the field.

Fairly recently a study was conducted in 2020 by Brian Sanders et al. [28]. They looked at the alternative method of controls, a gesture based system compared to a joystick, in this case a Xbox-controller. For this they conducted two different test, first a drone in a virtual environment, developed in Unity. The second test consisted of two parts. First they experienced and trained on a virtual car also developed in Unity, and later they used the same means of gesture control, but to control a real model car. The gesture input was achieved by using a Leap Motion controller in both tests. In the the first test were they controlled a drone some issues occurred with the LMC. First of all, fatigue was a problem for the gesture based system. The users made too big gestures, causing two issues, prematurely causing fatigue, but also going out of bound of the LMC field of view. This caused irritation with the participants because the inputs was not registered. These problems was not present with the Xbox-controller. To fix this some redesign took place, for example virtual version of the hands were displayed to show if they were in field of view or not. Support for the hands were also added. After this the test was successful and proved that the ability to control vehicle's with gesture-based controls is achievable. However, the participants prefered the Xbox-controller. For the second test they looked if training in a virtual environment could mean skills that are transferable to the real world. A model car was driven around in a world created in Unity, and the knowledge from the first test was brought in, so the same mistakes would not be repeated. After a training session, the participants controlled a real remote controlled car, and did so successfully. This showed that training in virtual reality works. So what they gather is that a gesture based interface could work, and can be intuitive.

2.5 Autonomous vehicles

The dream of autonomous driving is nothing new. Already back in the 30's, General Motors presented the idea of self driving vehicles at the Futurama exhibit [29]. But due to their high cost and lackluster demand the idea never caught on. Fast forward to October 2015, Tesla released Autopilot 7.0 [30] and with it the public got access to an autonomous driving that exceeded the conventional cruise control and adaptive cruise control. Nevertheless, there had been car makers before Tesla researching autonomous driving, but in a way Tesla is like Apple [31]. They brought auto pilot to the market at a reasonable cost, at that time, that not other car makers did.

There are five levels of automation levels according to SAE International [32], former Society of Automotive Engineers. Today these levels are classified as the standard of automation (see Figure 5). The "holy grail" of automation is level five, where the system will not require any attention of the driver and can drive in any condition everywhere. Level zero to two has been a common occurrence now in cars and ranges from adaptive cruise control and/or lane centering. Tesla is currently at level two (year 2020), and has started to beta test a level three were level three starts to be fully autonomous driving. The driver is not driving and it can handle traffic jams, but requires the driver to take over whenever the system requests it. Level four will not require the driver to drive at all, but it only works under limited conditions.

Furthermore, when looking at external Human-Machine Interface (eHMI) and its use case in autonomous vehicles, in a fairly recent study [33] done by Stefanie M.Faas and Martin Baumann, they looked at what color would be suitable as the standard light for indicating that a vehicle is in autonomous driving. They tested the color white and turquoise to see which one is more suitable. Almost everybody preferred turquoise over white (N=55 preferred turquoise over N=4 preferred white). Turquoise light is quickly and reliably identifiable as it does not interfere with any other colors that are present on the vehicle, and it is a novel color that is not associated with anything else yet.

Scania has developed a fully autonomous concept truck called AXL [34]. This is capable of fully autonomous driving (SAE level 4) and is suitable for environments such as mines or large closed construction sites. It is equipped with cameras, radar, GPS and light detection and ranging (LIDAR) sensors, all which together will create a good map of the trucks surrounding. Scania engineers said that this system would probably not be fit for a city, where the traffic and other hinders would be too unpredictable for the system. But in a mine or other closed environment where it is more predictable it would be more suitable. So according to SAE it would reach a level four of automation.

Volvo has developed a system called Volvo Dynamic Steering [35]. Besides creating effortless steering, there are some autonomous features with this system such as stability assist to prevent skidding and a lane keeping feature. They also have implemented an external steering. With the help of a joystick, the driver

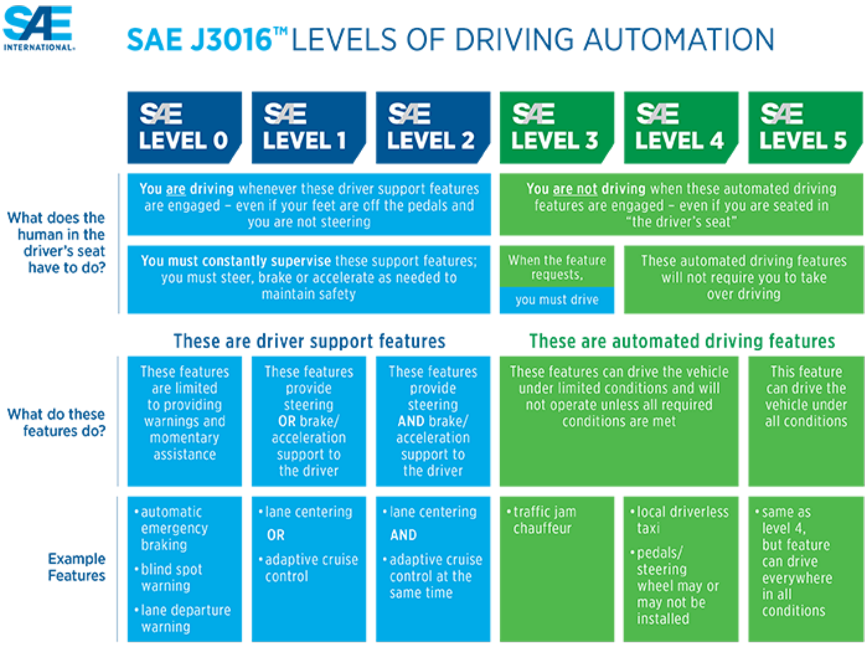

Fig. 5. The different automation levels, diagram found on SAE website [32].

can get out of the truck and drive the truck remotely while standing beside it. For safety, the truck has limitation of a top speed of 10km/h. In addition, there are other benefits with an autonomous vehicle fleet in mining environment. A thesis back in 2013 by Parreira, Juliana [36] showed that, productivity increase, improved fuel consumption's, less tire wear and more. Much of this could be credited to less stops as there is no need of stop signs in intersections, the autonomous fleet knows where every other vehicle is at all time.

3 Method

The most important aspect of this thesis was the prototype and the user testing. This section will explain all the methods used, the research, how the prototype was implemented and how the testing was conducted.

3.1 Litterature Review

To get a better understanding for this thesis a literature review was conducted to get an extensive view on how the market looks like and what is available. To do this we used Google Scholar and Umeå University's online database which has access to almost every scientifically article that is published. We also used Piteå city library that had some books that were not available for online reading at Umeå University.

Most of the research included:

- Gesture based interfaces for robots or vehicles
- Use of Kinect-device
- Autonomous vehicles
- Speech recognition
- Machine learning

3.2 Prototype Development

In this subsection we present the tools and how we used them to create our prototype.

3.2.1 Tools used The prototype was made in Unity version 2020.2.3f1, the camera used was the Kinect 2.0 and the firmware for Kinect was the official SDK version 2.0. Visual Gesture Builder, which is included in the Kinect offical SDK, was used for creating a gesture database which the prototype used for its gesture recognition. (See image 6 for Unity developing environment.) All of this will be developed on Windows 10.

3.2.2 Leap Motion Interface The first device we got access to was the leap motion controller (LMC). The offical SDK from Leap Motion was installed which also includes the drivers for the LMC. Sadly the app store at the moment did not work properly, so we were not able to try the LMC out of the box. However, the support for Unity was good and it was easy to start developing small programs for the controller to try it out. We did some smaller prototype for the LMC to get a better understanding of the use cases and how it works, but we realised that the narrow field of view was not suitable for the prototype that we had in mind. So the idea of using LMC was scrapped.

Fig. 6. Developing the prototype in Unity, not the finished version.

3.2.3 Microsoft Kinect Interface The hardware we chose for our final prototype was the Microsoft Kinect 2.0. This was paired with Unity, and the gesture database was generated with Visual Gesture Builder that is included with the official Kinect SDK for developers, where we tagged gesture and machine learning was used to train. To train the gesture, we used ourselves as models, but also the colleagues at Rise as well as friends and family. In total we had seven to nine people per gesture helping with recording video that was later used for tagging the gestures. None of these videos where shown to anyone else but us that made this prototype.

3.3 Set of gestures

For this research two set of gestures were constructed, see table 1 and 2. For these gestures we took inspiration from air marshalling, whose gestures are designed to be clear and universally understandable. But the air marshalling gestures are continuous(there is movement in the gesture) and for this prototype we needed static gestures, so some alterations had to be made.

3.3.1 Prototype Since it would be hard or impossible to get an actual working truck to test this hypothesis, a prototype of a virtual truck was made. This was so we would be able to test and evaluate if the use case of a gesture based interaction would be a suitable control for a driverless truck. For this research we

Table 1. Set of six gestures.

Action	Gesture
Forward	Both hands above head
Forward left	Right hand above head, left hand out to the left
Forward right	Left hand above head, right hand out to the right
Reverse	Both hands on the chest
Reverse left	Right hand on the chest, left hand straight out to the right
Reverse right	Left hand on the chest, right hand straight out to the left

Table 2. Set of four gestures.

Action	Gesture
Selected direction straight	Both hands above head
Selected direction left	Right hand above head, left hand out to the left
Selected direction right	Left hand above head, right hand out to the right
Toggle direction, forward or reverse	Both hands on chest

had both Kinect and Leap Motion at hand to use for the development, but ultimately we chose the Kinect as our device, since it was more suitable for gesture detection on a distance. As mentioned earlier, the tool we used for developing the prototype is Unity, version 2020.2.3f1. To display this we chose a 60" TV from the brand LG. We chose the largest TV we could get our hands on for the testing as we hypothesised that using larger screen would yield a greater immersion than a smaller computer screen. The reason we went with Unity is that it has a huge community and their assets store contains multiple free resources that would increase the development speed, and the need to model something our self would not be required. We did several iterations of the prototype to get it working. Some unofficial user tests were conducted to see how it would work when it was not us that used it, and then we changed it accordingly to get a working prototype.

3.4 Testing

As of the time this master thesis was conducted a pandemic was present (COVID-19). A lot of restrictions was in place which made it harder or impossible to conduct certain tests that would have been suitable and helpful for this research. Because of this we decided to not restrict which participants we chose. Preferably users that had some experience with heavy vehicles or wheel loaders were selected, but since this is also a futuristic idea there is no clear target group for this interface. Hence a variance in the participants background. Two pilot tests were also conducted before the full-scale test. This was so that if we missed any question, the prototype was not working as expected or any other unforeseen obstacle, no valuable test participants would be wasted.

To give the participants the same conditions before each test, a test script was printed that they where given. It included what this project was about, that it is the system that is tested, and not the participants and what the task was. The test was constructed so that the participants either tested the set of four or six gestures first, then the other. This was altered between each person, so if the first participant tested six gestures first, then the next participant tested four gestures first. This is to negate any potential bias against each other. To explain the gestures, two different set of papers were handed out that showed how the gesture looked like, and what action it corresponded. The supervisor was neutral during the test, if the participants got stuck with something, the leader would not help them. The only time the test leader was helping was if the participants would be stuck at the same time for far too long, and it would risk them not finishing their task.

The questions asked before the test was:

- What is your current occupation?
- Do you have any experience maneuvering heavy vehicles?
- What driver licences do you have?
- Have you previously used any gesture based interaction?
- What is your attitude towards a gesture based interaction system?

The object of the test was to back the truck in between the jersey barriers, during this test notes where taken on how they performed, how long it took and if the gestures gave them any hassle. The participants where also encouraged to speak out their thoughts aloud, this according to the "Thinking aloud protocol" [37]. Advantages of this method is that you can get preference and performance information directly, information that could be forgotten if asked after the test. Participants can reveal misconception and confusion in an early stage. Furthermore, to gather additional information the participants were asked questions straight after the test.

Those where:

- Did you prefer six or four gestures?
- What set of gestures were easier to learn?
- Of those two different set of gesture, which one did you prefer? And why?
- Were there any gesture that did not feel good?
- How was your experience with this prototype?
- What was the physical effort for the set of six gestures?
- What was the physical effort for the set of four gestures?
- Is this something you could potentially use in the future?
- Would you trust this system?
- Do you have any suggestions or other views?

Notes were also taken during the test, noting if the user was struggling with something or if there was any other issues. Test participant also signed a paper to give consent to the collection of data, according to GDPR.

3.4.1 Presentation To present the findings of this research two presentations was held. One seminar for Umeå University that will be held digitally on Zoom were the other students doing their master thesis at the same time, supervisors and friends and family will be present. The other will be for RISE, where project members in In The Hub from Scania, Boliden and ICE Makers also will be participant. This will be a more informal presentation.

4 Result

This chapter will present all the result that was obtained for this master thesis. This includes the result from literature review, the prototype and the result from the user testing.

4.1 Literature review

What we could find from other research is that this form of interaction models, where you control a heavy truck when standing next to it have not been extensively studied. However, there have been plenty of studies on gesture control on a variety of vehicles, such as UAV, RC car and robots, many of them were successful when controlling with gestures. However, one downside almost all of them had was the issue of fatigue, that prolonged static gestures were arm and hands are lacking support will cause strain. Furthermore there is the issue of precision with gesture control. If the task require delicate work, the lack of feedback and not having enough resolution will cause poorer accuracy. Nevertheless, gesture control could have potential to work if you find the right application for it. In addition, with the technology evolving, the application will become broader.

4.2 Prototype

The Hi-fi prototype created for this research was made in Unity version 2020.2.3f1. The program consist of two key aspects, the playable truck and its gesture detection. The finished prototype can be seen in the image 7. The Kinect was wired to the computer with a USB-cable. In order to know which direction the truck is headed, an arrow was placed on the truck, and changes accordingly to the chosen direction. One key aspect with gesture based interaction is correct feedback, but as we wanted as close to real life simulation, no image of the Kinect skeleton-tracking was implemented on the prototype.

4.3 User Test Result

All the test was conducted at the office of RISE Piteå. The test took part in late February to early March in 2021. There were 10 participants in total, six males and four females between 24 and 38 years of age. The result from the user test is presented below.

4.3.1 Gesture based interface Because of the low number of participants, no statistical significance could be extracted from the result. However, we still got data on how the participants felt about the interface. One key aspect that we could get from the result was that the majority of the participants perceived the gesture control as intuitive. When comparing the two sets of gestures, we did not get any clear view if one was favoured over the other. The result showed

Fig. 7. The prototype developed in Unity.

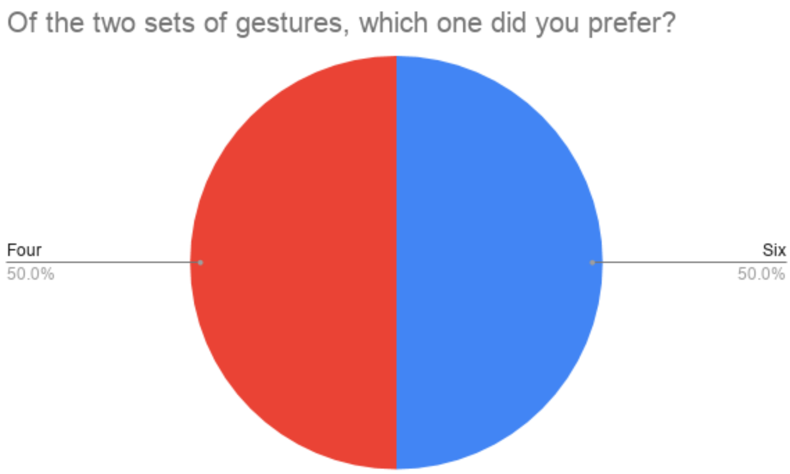

Fig. 8. Which set of gestures the participants preferred, where the number of participants is ten.

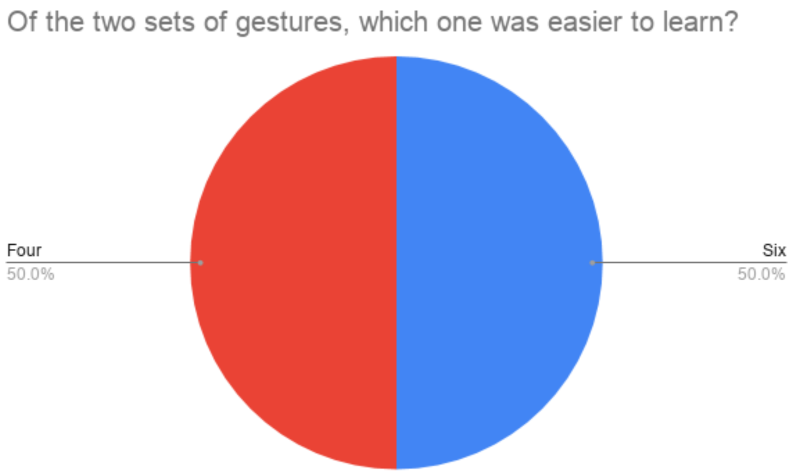

Fig. 9. Which set of gestures the participants thought were easier to learn, where the number of participants is ten.

a 50/50 split between the two of them. This was also the case in which one was easier to learn, see figure 8 and 9 for charts.

The next table, 3 has the different opinions the participants had about the set of gestures that they preferred. On the set of four gestures one response was missing due to misunderstanding.

Table 3. Opinions on why they favoured the gesture.

Six	Four
More obvious what direction the vehicle is going	Easier to remember
Easier to coordinate direction movement with turning	Felt more intuitive and fewer gestures to remember
Felt more natural	Fewer gestures to remember
Easier to remember the gestures	Fewer moments to remember initial
Easier with one gesture for each action	

Looking at figure 10 we have the answers from a multiple answers question. The participants were asked what they thought about the prototype. A majority of them leaned towards the positive side, with it being intuitive and useful. There were also those that thought that the prototype would not do as they wanted, that the gestures were not recognized properly.

Next we have table 4. This shows us the time it took for the participants to complete the task, either six gestures first then four, or vice versa. Looking at this data we can see that the average time to completion is lower for four gestures, both when the participants started with that gesture and when it was their

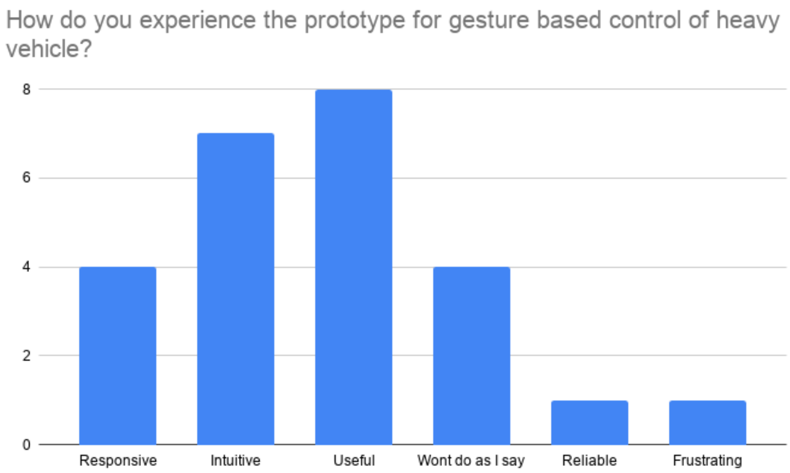

Fig. 10. Chart of how the participants felt about the prototype.

second turn. The participants performing the set of six gestures were generally slower than the participants performing the set of four gestures. One explanations for this is that the gesture detection was poorer in the set of six gestures.

Table 4. Time it took to complete the task, with either six or four gestures first. Time in minutes.

Six gestures, 1st test	Four gestures, 2nd test	Four gestures, 1st test	Six gestures, 2nd test
2.25	1.43	2.53	2.02
3.29	2.03	4.30	3.19
5.25	4.03	3.14	5.30
4.53	1.50	3.07	3.03
4.47	2.57	2.18	3.12
Average time to completion			
4.12	2.31	3.12	3.25

Regarding gesture recognition, a small part of the participants had issues with the system having low recognition for the gestures, with the backwards action causing most of the trouble. This caused frustration and the participants got insecure whether the gesture was recognized or the prototype worked at all. In addition, when a gesture did not work as intended, some thought that they were out of the Kinect's field of view, and moved a bit. They would also try to alter the gesture in some variation, and if that worked they thought that the initial gesture was wrong.

One of the more negative aspect was when the set of four gestures was used, participants complained about fatigue because the arms were always above the head, except when they changed direction. This problem was not as noticeable when the set of six gestures was in use. See figure 11 and 12 for the response.

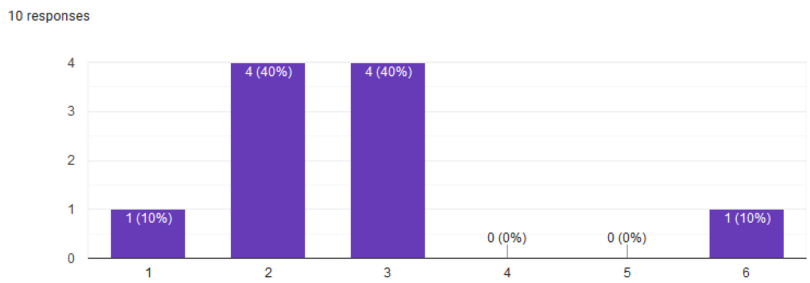

Fig. 11. On a scale of one to six where six is very heavy, how the participants experienced the level of fatigue when using the set of six gestures.

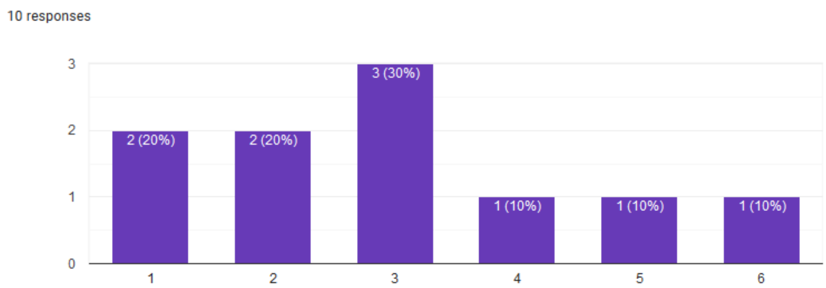

Fig. 12. On a scale of one to six where six is very heavy, how the participants experienced the level of fatigue when using the set of four gestures.

Looking at figure 13 we can see the responses on what the attitude is towards a gesture based interaction system. What we can see was that everyone is positive about it.

We also asked the participants if there were any other remaining viewpoints as a last question. One interesting points where brought up. When designing the gesture it would be wise to think it through, so that there are not any possibility for the gesture to be mistaken as something inappropriate.

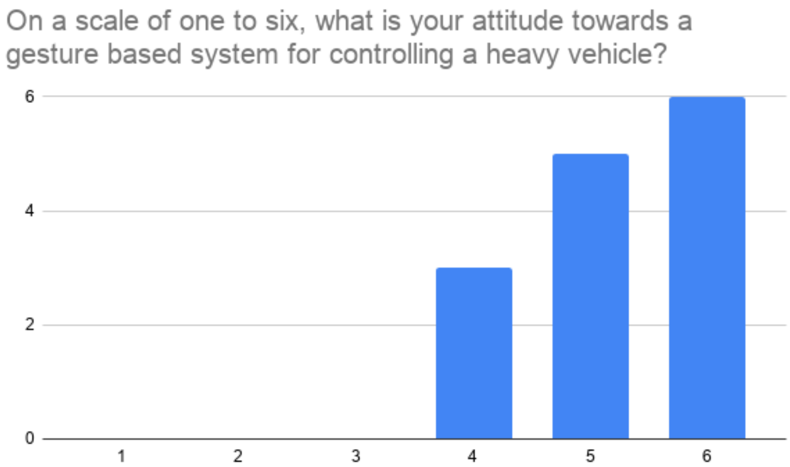

Fig. 13. On a scale of one to six were six is very positive, what the participants attitude is towards a gesture based system for controlling a heavy vehicle.

5 Discussion

In this chapter we will discuss our prototype and the user test result. Findings and our own opinions from the related research. The chapter will also cover further development and future work.

5.1 Related research

As David Procházka et al. [26] wrote, one of the key aspect is immediate feedback. If there are no feedback given straight away when performing a gesture, the user will be confused. This was confirmed by our own testing. If a gesture's performance was poor, the user started to doubt or think that they did something wrong. Some other issues with a gesture based control, especially camera based, is field of view in the sense that it can not be blocked. If the visibility is poor, either by smoke, bad weather or just plain out of sight, the gesture detection will work improperly or not at all. In addition, Brian Sanders et al. studie [28] proved that skills learned in a virtual environment can be transferred for real world applications. With this knowledge it is a possibility to use this kind of prototype, were participants learn and train how to control a heavy vehicle in a virtual environment first. Then use that skill in real world, saving both cost and time. It can also be looked at the way, if a proof of concept like this gets good feedback, a real life use would not be as far of as well.

Regarding alternative interaction models, we deem that speech recognition is not suitable for this kind of application, controlling a vehicle, in the sense of micro management should not be used with the help of speech. The load on the user will be far too heavy, as the commands has to be used repeatedly. This was noted with Komiya, Kayoko' et al. [23]. A better application for speech recognition would be like in RISE's unpublished work. For example, "Drive the truck to loading area", and then let the autonomous driving take over.

5.2 Prototype development

Discussion about what the thought process was when we developed the prototype, and how we solved some issues.

5.2.1 Gesture control When we set out to record these gestures for the recognition part we discovered that the first version of backward action, used both in the set of four and six, did not work properly. The first version the hands were on the chest, but this caused the skeleton joints bug out, and a high confidence level was hard to reach. To fix this all that was needed was to lower the hands a bit, to be just below the chest instead. After this the confidence level of the gesture recognition was high for all the gestures when testing on ourselves. Regarding the issues where some of the participants had problems with gestures not being recognized it is mainly because of the machine learning. From what we learned the gestures should be trained into several different gestures, that executes the same action. For example two participants would raise their hands over their head in two different ways, even if the change is subtle. This way we would overcome the problems with over training a gesture, but also get better confidence for a wider range of users. Also worth noting is that this problem was something that occurred with the Visual Gesture Builder. The same problem may not happen if the gestures are trained in a different way.

Furthermore, even if we did not get a set of gestures that were clearly preferred, we still got acknowledgement that we are on the right track. They are intuitive, and with some more iterations to solve the fatigue issues they would be more suitable for use.

5.2.2 Prototype When designing the prototype we chose not to do a lo-fi prototype. We needed all the time we could get to develop the hi-fi prototype, and we quickly realized that the development speed in unity was fast enough that there were no need. So the prototype went through some iterations. First we made a playable prototype with a keyboard with a simple environment. After we got the Kinect-camera we implemented a basic control just to verify that it worked. After that more gestures were integrated until we got a working prototype. However, one of the mishaps with the prototype that we discovered far too late was that it would not build. We found that one of the script was causing the issue, but there was not enough time to fix it, since the testing was scheduled in a few days. So to work around this the prototype was optimized in the way that non-necessary props was removed from the simulator, and we ran it inside Unity development tool. There were no noticeable performance issues and the tests could be conducted. Furthermore, we initially developed the prototype to be used with a virtual reality headset, Oculus Rift, for a greater immersive experience. But since it is hard to sanitize the fabric of the headset, a large flat screen TV was used. Another improvement the prototype could get is to make it more fluid. As it is now, when going straight from forward to forward left, the truck breaks since it does not recognize a gesture between those to and so it jitters. But if there would be a check to see if the gesture is changing to another and not hit the breaks, the whole experience would be more fluid and pleasant. For further research we could look into the Leap Motion controller again. Since only the hands are required the gestures don't have to be as large, and in such case the fatigue could be minimized.

5.3 Test result

As mentioned in the earlier chapter, the test result that we gather could not be used to get any statistical significance. Too few participants were tested. However, the test still showed promising result and can act as guidelines for further research.

From what we could see a gesture based interaction model, as in this case, controlling a vehicle could potentially work. A large part of the participants felt that the interaction model for controlling the truck in the prototype was intuitive, and was surprised how easy and natural it felt. Nevertheless, there are some issues that would have to be solved. The first important topic to discuss is that some participants struggled with gestures not being recognized. There were one in particular that caused most of the issues, the reveres/toggle direction action. This gesture was for some of the participants hard to activate, and thus causing frustration and performance issues. Since a smooth experience heavily relies on the gestures to work properly, this is a key aspect that has to work. In the testing we saw that if some gestures were not working, it did not take long before the users start to doubt if the gesture control works. The difference in time to complete the task could also be traced back to problems with poor gesture detection. We saw that the set of four gestures got an average better time than the set of six gestures. The main reason behind this is that since the set of gestures of four utilized less the gesture that activated reverse/toggled direction, the time was better compared to the set of six gestures. Additionally, when designing the gestures the ergonomics has to be taken into consideration. One thing that was noticed during the test was that the set of four gestures caused more fatigue on the participants than the set of six gesture. This was because of how the gestures was designed and almost all of the gestures where with the hands above shoulder level.

In addition, because of the fixed point of view in the camera, the angle that was chosen to display the virtual environment had some issues. Mainly the perspective and knowing which way to turn while reversing. Since the camera angle was a fixed point, it was hard to see how far or near the participants where to the jersey barriers. Many of the wrong turns was also caused by this. Since usually you sit inside the vehicle when backing up, turning the wheel is natural when reversing. This was different when standing outside and looking in front of the vehicle. Now the participants had the think the opposite to turn correctly, causing confusion.

Another aspects of these test were that they got no training, if it had been given, we would be looking at completely different result. Mainly the participant would not guess if they were doing the gesture correct, since their only reference on the actual test was a picture given to them, depicting the different gestures with the action according.

From what we gather, the test showed promising result. But if gesture based interfaces are going to be used, the rate of error has to be extremely low. Users get frustrated quickly if the interface does not respond fast enough, either due to bad performance or if the gesture is not recognized. Also the gesture recognition is the key aspect that focus has to lie on. With today's hardware, performance issue caused by poor performing computers is rare. In addition, to negate bad gesture recognition, the gesture itself could be really simple, or a huge amount of time in Visual Gesture Builder has to be spent tagging gestures and getting it right for everyone.

5.4 Future work

The work we did is not complete in any way. There is still a long way to go, but from the prototype and testing we conducted, we could see that there are potential in this field. Gesture based interaction models is something that could be useful in the future, and in this case we used the Kinect to get a skeleton tracking, and used that for gesture recognition. But with Googles Mediapipe Holistic, any camera could potentially be used for gesture recognition. That would mean that even existing cameras could potentially get upgraded with new firmware, and therefore be able to recognize gestures. Even tough we chose not to develop and test the Leap Motion Controller, this is something that could have potential. Since our initial idea was focused when the operator was standing next to outside of the vehicle, we deemed the LMC not suitable. However, for other use cases this could be an interesting idea to explore. The LMC would not require any gesture to be above the head, so the risk of fatigue would be less. Moreover the company that makes the LMC is further developing their controller, some even with haptic feedback. Haptic feedback could be used to make sure that the user is inside the controllers field of view, and in such case negating the problem with users going outside of the controllers FOV.

In addition, one thing that we did not look at all but still is very important is the access to the gesture control. Who shall be allowed to control a vehicle, and how would that security check look like. Face recognition, special equipment being on the operator is some different options that could be explored.

5.5 Further development

This research served as a pre-study for SCANIA, RISE, Boliden and Icemakers. They will continue to develop a prototype for use case A, and a variant of the prototype that was tested in this master thesis will be further developed. A gesture based interface has showed that it could potentially work and would require further research and testing.

5.6 Restriction & Limitations

As mentioned earlier, during this master thesis a pandemic was present, Covid-19. This caused issues in many ways as the recommendations from the public health agency of Sweden, Folkhälsomyndigheten, urged the citizen to keep a social distance and work from home as much as possible. Because of this we could not do as much user testing as we liked, and it also made it inconvenient to do study visit in for example mines and other work areas where heavy vehicles are present. The master thesis also covers 30hp which corresponds to 20 weeks of full time study, so the time for this research is limited.

6 Conclusion

With the rise in autonomous vehicles the need for intuitive system to interact is growing. This thesis aimed to develop and test if a gesture based interaction model to control an autonomous truck when standing outside, next to it, is a suitable mean of control. It looks at what the interaction model needs to make it reliable, trusted and usable. We successfully made an interaction model for controlling an autonomous vehicle with gestures, which could also be seen as a great result. Almost all of the participants felt that a gesture based interaction is intuitive, and the interaction felt natural for the participants. There were no major difference between the two set of gestures. Both was on same level of intuitiveness, with one minor difference that the set of six gestures put less trains on shoulders and arms. Knowing this the set of gestures would require another iteration, to both get a higher recognition level in addition to removing the straining issue on shoulders and arms caused by lifting the arms above shoulder level. Furthermore, most of the participants felt that they could trust this form of interaction, but it needs to mature and have heavy safety feature, for example like the one that was implemented in the prototype. If no gestures are detected, the truck will brake and therefore prevent it from running away.

In conclusion, a gesture based interaction model for autonomous vehicles could potentially work, but the gesture recognition has perform at a level that guarantees a high level of confidence in recognition to prevent frustration for the user. This would be achieved with either simple gestures, or a well trained database with gestures. The gestures would also be design in a way that does not cause fatigue on the user.

Furthermore, when looking at the previous research done by others and compare it to our testing and result, shows that gesture based interactions keeps improving, as a result of better hardware and better understanding of use cases.

7 Acknowledgements

I would like to thank my supervisor at Umeå University, for the support during this master thesis. Second, I would like to thank my supervisor at RISE that has helped me throughout the project, and welcoming me to their team. Also thank you to the peer reviewers for their feedback. Lastly I would like to honor my beloved father who recently passed away. Thank you for always supporting me.

Thank you.

Bibliography

[1] D. F. Armstrong and S. Wilcox, *The gestural origin of language*, ser. Perspectives on deafness The gestural origin of language. Oxford: Oxford University Press, 2007.

[2] C. Duhigg, *The power of habit : why we do what we do and how to change*, 2014th ed. New York: Random House Trade Paperbacks, 2014.

[3] "In the hub," 2020, https://www.ri.se/en/what-we-do/projects/hub, accessed 2020-11-16.

[4] R. Bolt, ""put-that-there": Voice and gesture at the graphics interface," pp. 262–270, 1980. [Online]. Available: https://doi.org/10.1145/800250.807503

[5] J. Lowensohn, "Kinect's launch day bumps and triumphs," 2010, https://www.cnet.com/news/kinects-launch-day-bumps-and-triumphs/, accessed 2020-09-29.

[6] "Leap motion has launched," 2013, https://blog.leapmotion.com/leap-motion-has-launched/, accessed 2020-09-30.

[7] J. Bakalar, "Nintendo wii (original, wii sports bundle) review," 2006, https://www.cnet.com/reviews/nintendo-wii-original-wii-sports-bundle-review/, accessed 2021-03-17.

[8] J. Lien and N. Gillian, "Soli radar-based perception and interaction in pixel 4," 2020, https://ai.googleblog.com/2020/03/soli-radar-based-perception-and.html, accessed 2021-03-17.

[9] "Visual gesture builder: Overview," 2014, https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn785529(v=ieb.10)?redirectedfrom=MSDN, accessed 2020-11-19.

[10] J. Corde, "Farewell, dear sweet kinect," 2018, https://www.windowscentral.com/ode-kinect, accessed 2021-03-17.

[11] M. Wilson, "Exclusive: Microsoft has stopped manufacturing the kinect," 2017, https://www.fastcompany.com/90147868/exclusive-microsoft-has-stopped-manufacturing-the-kinect, accessed 2021-03-17.

[12] J. Jiao, L. Yuan, W. Tang, Z. Deng, and Q. Wu, "A post-rectification approach of depth images of kinect v2 for 3d reconstruction of indoor scenes," *ISPRS International Journal of Geo-Information*, vol. 6, p. 349, 11 2017.

[13] "Kinect on xbox one will not record or upload your conversations, microsoft says," 2013, https://www.polygon.com/2013/6/6/4403926/kinect-on-xbox-one-will-not-record-or-upload-your-conversations, accessed 2021-12-01.

[14] "What are the specifications of the leap motion controller?" 2019, https://support.leapmotion.com/hc/en-us/articles/360004476658-What-are-the-specifications-of-the-Leap-Motion-Controller-/, accessed 2020-11-16.

[15] L. E. Potter, J. Araullo, and L. Carter, "The leap motion controller: A view on sign language," in *Proceedings of the 25th Australian*

Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, ser. OzCHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 175–178. [Online]. Available: https://doi-org.proxy.ub.umu.se/10.1145/2541016.2541072

[16] "Leap motion controller," 2021, https://www.ultraleap.com/product/leap-motion-controller/, accessed 2021-12-01.

[17] "Adaboosttrigger," 2021, https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn785522(v=ieb.10), accessed 2021-03-18.

[18] "Unity platform," 2021, https://unity.com/products/unity-platform, accessed 2021-03-18.

[19] "Scripting," 2021, https://docs.unity3d.com/Manual/ScriptingSection.html, accessed 2021-03-18.

[20] "Mediapipe holistic," 2021, https://google.github.io/mediapipe/solutions/holistic.html, accessed 2021-03-18.

[21] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[22] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[23] K. Komiya, K. Morita, K. Kagekawa, and K. Kurosu, "Guidance of a wheelchair by voice," in *2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies*, vol. 1, 2000, pp. 102–107 vol.1.

[24] "Aircraft marshalling and ramp hand signals - elearning," 2021, https://www.iata.org/en/training/courses/aircraft-marshalling-ramp/talp58/en/, accessed 2021-03-18.

[25] "Creative air marshalling," 2014, https://fearoflanding.com/demystifying/creative-air-marshalling/, accessed 2021-11-04.

[26] D. Procházka, J. Landa, T. Koubek, and V. Ondroušek, "Mainstreaming gesture based interfaces," *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, vol. 61, no. 7, pp. 2655–2660, Dec. 2013. [Online]. Available: https://doi.org/10.11118/actaun201361072655

[27] S. G. B. M. Elliott, Linda R. Hill, "Gesture-based controls for robots: Overview and implications for use by soldiers," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2016. [Online]. Available: https://apps.dtic.mil/sti/citations/AD1011904

[28] B. Sanders, Y. Shen, and D. Vincenzi, "Design and validation of a unity-based simulation to investigate gesture based control of semi-autonomous vehicles," in *Virtual, Augmented and Mixed Reality. Design and Interaction*, J. Y. C. Chen and G. Fragomeni, Eds. Cham: Springer International Publishing, 2020, pp. 325–345.

[29] A. Townsend, "The 100-year history of self-driving cars," 2020, https://onezero.medium.com/the-100-year-history-of-self-driving-vehicles-10b8546a3318, accessed 2021-03-21.

[30] "Your autopilot has arrived," 2015, https://www.tesla.com/blog/your-autopilot-has-arrived?redirect=no, accessed 2020-12-09.

[31] V. Savov, "The 100-year history of self-driving cars," 2017, https://www.theverge.com/2017/7/31/16067728/tesla-model-3-apple-iphone-comparison, accessed 2021-03-21.

[32] "Sae international releases updated visual chart for its "levels of driving automation" standard for self-driving vehicles," 2018, https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-"levels-of-driving-automation"-standard-for-se accessed 2020-12-10.

[33] S. M. Faas and M. Baumann, "Light-based external human machine interface: Color evaluation for self-driving vehicle and pedestrian interaction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 1232–1236, 2019. [Online]. Available: https://doi.org/10.1177/1071181319631049

[34] "A new cabless concept – revealing scania axl," 2019, https://www.scania.com/group/en/home/newsroom/news/2019/a-new-cabless-concept-revealing-scania-axl.html, accessed 2020-12-11.

[35] "Volvo dynamic steering on volvo fh, safety to the next level," 2021, https://www.volvotrucks.com/en-la/trucks/volvo-fh/features/volvo-dynamic-steering.html, accessed 2021-03-18.

[36] J. Parreira, "An interactive simulation model to compare an autonomous haulage truck system with a manually-operated system," Ph.D. dissertation, University of British Columbia, 2013. [Online]. Available: https://open.library.ubc.ca/collections/ubctheses/24/items/1.0074111

[37] J. Rubin and D. Chisnell, *Handbook of usability testing how to plan, design, and conduct effective tests*. Wiley Pub., 2008.

[38] P. Daukintis, "Visual gesture builder – kinect 4 windows v2," 2014, https://peted.azurewebsites.net/visual-gesture-builder-kinect-4-windows-v2/, accessed 2021-03-17.

[39] "Sae internationl," 2021, https://www.sae.org/, accessed 2021-03-17.

[40] J. M. Szymański, J. Sobecki, and P. Chynał, "Actiontracking in gesture based information systems," in *Proceedings of the 2014 Mulitmedia, Interaction, Design and Innovation International Conference on Multimedia, Interaction, Design and Innovation*, ser. MIDI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1–7. [Online]. Available: https://doi-org.proxy.ub.umu.se/10.1145/2643572.2643590

[41] L. Zhao, "Gesture control technology: An investigation on the potential use in higher education," 2016. [Online]. Available: https://blog.bham.ac.uk/itinnovation/2016/03/25/gesture-control-technology-an-investigation-on-the-potential-use-in-higher-education/