THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Datadriven Human Intention Analysis

Supported by Virtual Reality and Eye Tracking

Julius Pettersson

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2021

Datadriven Human Intention Analysis
Supported by Virtual Reality and Eye Tracking

Julius Pettersson

This thesis has been prepared using LaTeX.

Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Reproservice
Gothenburg, Sweden, June 2021

To family and friends, eye thank you.

Abstract

The ability to determine an upcoming action or what decision a human is about to take, can be useful in multiple areas, for example in manufacturing where humans working with collaborative robots, where knowing the intent of the operator could provide the robot with important information to help it navigate more safely. Another field that could benefit from a system that provides information regarding human intentions is the field of psychological testing where such a system could be used as a platform for new research or be one way to provide information in the diagnostic process. The work presented in this thesis investigates the potential use of virtual reality as a safe, customizable environment to collect gaze and movement data, eye tracking as the non-invasive system input that gives insight into the human mind, and deep machine learning as the tool that analyzes the data. The thesis defines an experimental procedure that can be used to construct a virtual reality based testing system that gathers gaze and movement data, carries out a test study to gather data from human participants, and implements an artificial neural network in order to analyze human behaviour. This is followed by four studies that gives evidence to the decisions that were made in the experimental procedure and shows the potential uses of such a system.

Keywords: Virtual reality (VR), time series analysis, human intention prediction, eye tracking, deep machine learning, uncertainty estimation, collaborative robots, psychological testing.

List of Publications

This thesis is based on the following publications:

[A] **Julius Pettersson** and Petter Falkman, "Human Movement Direction Classification using Virtual Reality and Eye Tracking". *Published in Procedia Manufacturing, Volume 51*, (pp. 95-102), 2020.

[B] **Julius Pettersson** and Petter Falkman, "Human Movement Direction Prediction using Virtual Reality and Eye Tracking". *In 2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, virtual, 2021.

[C] **Julius Pettersson** and Petter Falkman, "Human Arm Movement Intention Prediction using Eye Tracking". Submitted to IEEE Transactions on Industrial Informatics.

[D] **Julius Pettersson**, Anton Albo, Johan Eriksson, Patrik Larsson, Kerstin W. Falkman, and Petter Falkman, "Cognitive Ability Evaluation using Virtual Reality and Eye Tracking". *In 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, (pp. 1-6), 2018.

Other publications by the author, not included in this thesis, are:

[E] Dahl, M., Albo, A., Eriksson, J., **Pettersson, J.**, and Falkman P., "Virtual Reality Commissioning in Production Systems Preparation". *In 2017 22nd IEEE International Conference on Emerging Technologies and Automation (ETFA)*, (pp. 1-7), 2017.

Acronyms

VR:	Virtual Reality
VRE:	Virtual Reality Environment
HMD:	Head Mounted Display
ET:	Eye Tracking
UE:	Uncertainty Estimation
AI:	Artifical Intelligence
ML:	Machine Learning
DML:	Deep Machine Learning
SML:	Supervised Machine Learning
ANN:	Artifical Neural Networks
FNN:	Feedforward Neural Networks
CNN:	Convolutional Neural Networks
RNN:	Recurrent Neural Networks
RPM:	Raven's Progressive Matrices

Contents

II Papers 65

Part I

Overview

CHAPTER 1

Introduction

The ability to determine what actions or decisions a human is about to make can be useful in multiple areas, for example, in manufacturing where humans working with collaborative robots is becoming increasingly more popular [1]. The advantages of having humans and robots in the same workspace interacting with each other are many, such as; increased flexibility [2] and increased productivity for complex tasks [2]. However, the robots are still not that interactive since they cannot yet interpret humans and adapt to their swift changes in behaviour in a way that another human would do. The main reason is that the collaborative robots today are limited in their sensory input, which makes it the responsibility of the human to stay out of the way. Human intention prediction can be achieved using camera images and probabilistic state machines [3] with the goal of determining between explicit and implicit intent. Other ways are to monitor the gaze to predict an upcoming decision [4], analyze bioelectric signals, such as electromyography, to predict human motion [5], or use a combination of eye gaze and movement tracking to predict the goal location of a movement [6].

Another field that could benefit from a system that provides information regarding human intentions is the field of psychological testing. Testing of mental capacity has been around since the early 1900s and has been greatly extended since then [7]. These tests can be used, for example, to evaluate special abilities, intelligence and social attributes as described in [7]. There is, however, potential to improve these methods even further using the technology that is at hand today. It is often, during certain tests where the participant is asked to complete a specific task, as important to observe the person's behavior during the experiment as to obtain the actual test results [8]. The authors of [8] further describe that the test results will be affected if the person taking the test is anxious, showing signs of speech or language

difficulties or has difficulties concentrating.

Other fields that have been rapidly expanding and that may be used to provide an understanding of human behaviours and intentions are; virtual reality (VR), eye tracking (ET), gathering and management of large datasets, and artificial intelligence (AI).

A way to gather more insight into how a person is reasoning is to measure and analyze where the person is looking [9] and the technique of doing this is called ET. It is, for example, possible to gain insight into what alternatives the person is considering or what search strategy is used while performing a task [10], based on what a person is looking at. ET has, for example, been used in an industrial context to; use the gaze as the input for machine control [11], analyze industrial visualization of information [12], and evaluate new ways to facilitate human–robot communication [13]. It has also been used as a tool to diagnose autism [14], where children are participating in different games and social activities on a tablet while their gaze is being observed.

VR can be described as a technology through which visual, audible and haptic stimuli is able to give the user a real world experience in a virtual environment [15]. Benefits, such as being able to provide more relevant content and present it in a suitable context [16], are reasons to promote the use of VR in neuropsychology. It is suggested by [16] that the use of VR makes it possible to measure data such as accuracy, timing and consistency and, with that, improve the post process analysis. A virtual classroom has previously been implemented in a VR-environment to aid the assessment process of children with attention deficit hyperactivity disorder (ADHD) [17]. It can also be used in an industrial context; when making prototypes [18], to train operators in assembly [19], and improve remote maintenance [20].

The use of modern technologies such as ET and VR makes it possible to collect larger amounts of data, with higher accuracy, and at a higher pace than before [21]. These large volumes of data, created at high speed, and with great variety [22] is referred to as Big Data. One area of AI that can be used to process these huge datasets is called deep machine learning [23]. Big data and AI has been shown to be important tools for the future to improve industrial manufacturing [24]–[26] as well as providing benefits in the field of psychology, for example, when analyzing how students perform on cognitive diagnostic assessments [27] and to determine if a person has attention deficit hyperactivity disorder (ADHD) [28].

1.1 Research Questions:

The ability to determine an upcoming action or what decision a human is about to take, can be useful in multiple areas, for example, in manufacturing where humans are working with collaborative robots and in psychological testing where such a system could be used as a platform for new research or be one way to provide information in the diagnostic process. A way to gather more insight into how a person is reasoning is to measure and analyze where the person is looking [9] using ET. There are multiple ways of tracking gaze and one of them is through VR. The data that is collected needs to be analyzed and one area of AI that can be used to process these datasets is called deep machine learning [23] (DML). This is the basis for the following research questions:

RQ1: *Is is possible to predict human intention through the study of eye gaze?*

The importance of understanding human intention is becoming more and more important as described earlier. There are several ways of achieving this, for example, using camera images, electromyography, or a combination of eye gaze and movement tracking. The eye gaze can reveal what alternatives a person is considering or what search strategy is used while performing a task. The goal of RQ1 is to investigate if it is possible to predict human intention through the study of eye gaze.

RQ2: *Is DML a suitable tool to analyze the connection between eye gaze and intention in humans?*

DML has shown to be a powerful tool to analyze large amount of complex data in multiple research fields, including industrial applications and psychological research. RQ2 aims at exploring if the combination of eye tracking and DML could be used as a flexible tool to analyze the connection between eye gaze and human intention as different tasks are being carried out.

RQ3: *How can a VRE-test be designed to gather the necessary eye gaze and movement data to be used in an application based on DML?*

VR has successfully been used in both industry and psychological research. The benefits of using a VRE includes, for example, being able to continuously gather data from both the user and the environment, and it gives the developer of the test full control over all events in the VRE while providing the user with an experience that is similar to a real world application. RQ3, therefore, aims at determining a procedure for how a VRE can be designed to gather eye gaze and movement data, from human participants, that can be analyzed using DML.

1.2 Thesis Outline

The thesis is divided into two parts, Part I features an overview of the research and Part II contains the publications that constitute the basis of the first part. Part I starts off in Chapter 1 with an introduction to the field that has been researched, followed by theoretical introductions to the areas of psychological testing, virtual reality (VR), eye tracking (ET), and supervised machine learning (ML) in Chapter 2, Chapter 3, Chapter 4, and Chapter 5 respectively. The experimental procedure of using VR, ET, and ML as the solution to a data-driven problem is introduced in Chapter 6 followed by Chapter 7 that covers the presentation of four studies of prediction of human intention. The summary of the included papers is given in Chapter 8 and the thesis ends with a discussion, concluding remarks, and suggestions of future work in Chapter 9.

CHAPTER 2

Psychological Testing

In [29] the aim of a psychological test is described as a method to measure different abilities and conditions that cannot be directly observed, such as intelligence, psychopathology or neuropsychological disorders. Psychological tests are often standardized to ensure validity and reliability.

A psychological test is usually designed with a particular population in mind. An individual's result on the test is always presented in relation to this population, on an appropriate scale, for example IQ in cases where intelligence is measured. In a process called standardization, the test is used with a representative sample of the population [8]. From this group's mean values and variance, you then generate a function from raw points to the desired scale.

The reliability and validity of the test, i.e. if the same results are achieved as the measurements are performed multiple times and how well it measures what it intends to measure [8], also has to be calculated. One way to ensure reliability is to standardize the test procedure, for example making sure that the instructions given to the test person are always the same and that the environment in which the test is performed is the same [8], i.e. there is no external interference.

Another key element is to inform the participants about the premise of the testing and what their information will be used for to make them feel comfortable before giving their consent to participate [8]. There are additional factors, described by [8], that might affect the test results and/or the behaviour of the individual being tested such as anxiety, difficulties to concentrate or to communicate.

When collecting data for psychological research through the use of psychological testing this is mostly done manually. This means that researchers are often limited in the amount

and types of data that can be collected. Observations of behaviour are, for example, made in real-time or through watching video recordings [30] of the test participant.

2.1 Digitalization in the Field of Psychology

Virtual Reality (VR) is a technology that has proved useful within psychology, for example as a tool to observe the level of distraction amongst children with ADHD [31], [32]. The authors of [16] highlights the possibilities and benefits of measuring data using VR, such as accuracy, timing and consistency to enhance the analysis. The research in [33] shows that VR can be used to interact with children through facial emotions and expressions. It can also be of great use in the process of treating and rehabilitating arachnophobia [34].

Another field of technology, that is already part of psychological research today, is the study of eye gaze movement. The eyes contain multiple levels of information, for the sender as well as the receiver, about the environment, emotional and mental states [35]. Assessing eye movement through ET is already widely used today. It is, for example, used for research purposes, in areas such as theory of mind [36] (the ability to imagine other peoples feelings and perspective), diagnosing autism [14], as an assistive tool for people with mobility difficulties [37], evaluating responsiveness to joint attention in infants, as well as in diagnosing Williams syndrome, ADHD, and reading disabilities [38]–[40].

Previous research has also shown that ML has potential within psychology to predict and increase our understanding of behaviour [41]. Furthermore, a study has shown that ML is efficient in facial recognition to determine facial expressions [42]. Consequently this could provide another parameter towards the purposes of analyzing an individual's behaviour since facial expressions are closely tied to emotion [43]. Another study by [27] shows that both supervised and unsupervised artificial neural networks (ANNs) can be used to analyze how students perform on cognitive diagnostic assessments. It has also been shown in [28] that ANNs can be used to determine if a person has attention deficit hyperactivity disorder (ADHD) and results by [44] indicate that ML can be used for automated test scoring of a novel story recall task.

2.2 Raven's Progressive Matrices

The concept of general cognitive ability, the g factor, was first introduced by the English psychologist Charles Spearman in 1904 [45]. To distinguish the differences between general intelligence and specific abilities while performing different tasks, [45] also states a second factor named s. The g factor has two main components; the capacity to think clearly and make sense of complex data, called educative ability, as well as the capacity to store and reproduce information, called reproductive ability [46]. The s factor is often represented by a circle with four elements; spatial, logical, mechanical and arithmetical abilities [47].

Raven's Progressive Matrices (RPM) are a set of tests designed to measure abstract reasoning and g factor [48]. They are well known and widely used since they are easy to administer and to interpret in a clear way [49]. The RPM are graphically easy to implement in a virtual environment, and are thus well suited to implement in VR. These tests are available in three different forms; Standard Progressive Matrices (SPM), Colored Progressive

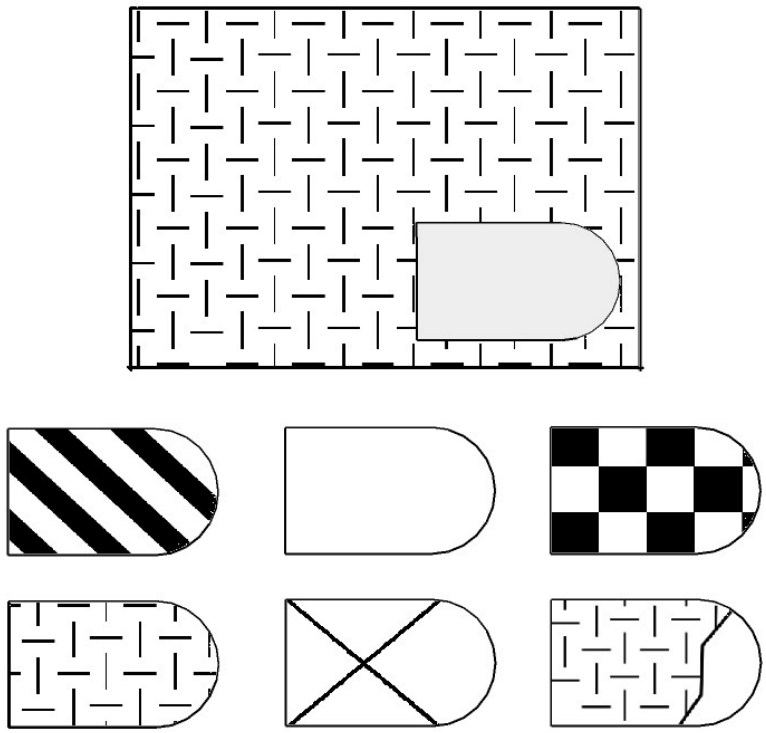


Figure 2.1: A figure showing an example item from Raven's standard progressive matrices.

Matrices (CPM) and Advanced Progressive Matrices (APM). These different versions are intended to be used for testing people with varying cognitive and physical abilities where SPM is the most widely used and was intended to be used once the intellectual capacity to reason has developed, age 8 and above. The CPM, on the other hand, was designed to be used before this ability has developed [50], age 5-11, and the APM was developed to be used on adults and adolescents with over-average intelligence.

The SPM test consists of 60 items divided into 5 sets (A-E) of increasing difficulty and was first published in 1938 [46]. Each set follows a different logic that progressively increases in difficulty [51] with each set becoming more difficult than the previous. Each item has a logical pattern where one piece is missing and the task is to select the correct alternative amongst a given set of alternatives, which varies from six to eight depending on the item and level of difficulty. An example from SPM of how these items may look can be seen in Figure 2.1.

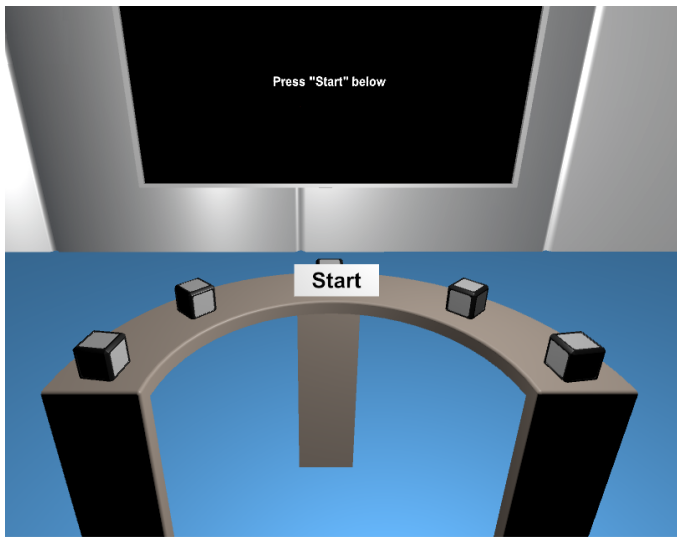
CHAPTER 3

Virtual Reality

Virtual Reality (VR) is the technique of 3D immersion in a computer created environment. A device that can be used to visualize the VR environment (VRE) to the user is a head mounted display (HMD) [15]. The HMD is, according to [15], equipped with sensors that measure the user's head motions and a display that is responsible for providing the user with the visual content. The system also provides the user with audible and haptic stimuli to immerse the user in a real world experience [15] of the VRE. An example of a person wearing an HMD can be seen in Figure 3.1a and the users view of the VRE from **Paper A** is shown in 3.1a. There are other ways that can be used to visualize a VRE, e.g. CAVE [52] that projects images on the walls of a physical room.

VR technology is spreading to new areas with a steady increase in overall usage [53], for example, in the field of psychological testing [17]. It has been used to measure the distraction level of children with attention deficit hyperactivity disorder (ADHD) [31], [32], in a virtual classroom. In [33], it is shown how VR can be used to imitate emotions and facial expressions for children to interact with. Recent research has proven VR to be useful in treating arachnophobia [34]. Benefits such as being able to provide more relevant content and present it in a suitable context [16] are other reasons to promote the use of VR. It can also be used in other areas, for example; when making prototypes [18], to train operators in assembly [19], and improve remote maintenance [20].

(a) A figure showing a person wearing a VR-headset consisting of an HMD and two hand-held controllers.



(b) The users view of the VRE from **Paper A**.

Figure 3.1: An example of a VR-headset and the view from inside the HMD.

CHAPTER 4

Eye Tracking

Eye tracking (ET) is defined by [10] as the technique of measuring what a person is looking at, in what order the objects are gazed upon and for how long the eye gaze stays fixed on that object. The eye gaze is an interesting biological marker because it is possible to analyze underlying neurophysiology based on the movement of the eyes [54]. Tracking gaze is therefore an appealing test method and also because it is objective, painless, and noninvasive [54]. ET can give an insight into the individual's problem solving, reasoning, and search strategies [10]. However, ET is only capable of tracking visible movements of the eye and not the hidden mental processes of visual attention [55]. This makes for the simplified assumption, when using ET for attention analysis, that attention is associated with gaze direction [55] even though that is not always the case.

One way of tracking the eyes, as described in [10], is achieved by illuminating them with infrared light, which is used to prevent the user from being dazzled, to get a clear reflection that is captured using a camera. The reflections are then used to calculate a vector of the relationship between the cornea and pupil [10], which in turn is used to calculate the gaze direction.

ET has, for example, been used to analyze the navigational intent in humans and how they interact with autonomous forklifts [56], analyze the prospective memory, used for delayed intentions, in children [57], investigate pedestrians' understanding of an autonomous vehicle's intention to stop at a simulated road crossing [58], to allow people with severe speech and motor impairments to move a robotic arm [59], and to predict which one out of four tasks, where the participants aligned two cubes in various ways in a VRE, that was carried out [60].

4.1 Eye Movements

The way we humans react to a visual stimuli is dependent on many factors [61], e.g. for a simple task we may be interested in determining if something is present or what it is, called detection and identification respectively, and for a more complex situation the goal might be to detect a target in a larger visual field of many targets.

In order for us humans to observe an object in the real world, we have to fixate our gaze at it for long enough time so that the brain's visual system is able to perceive it [62]. We are only able to see a very narrow visual scene with high acuity at any point in time [62] and to observe a larger area with acuity we need to continuously scan it with small rapid movements so called saccades. The fovea is a small area on the retina that is responsible for providing this high-acuity vision [62] using the lens that focuses the light coming from the pupil on this area that is densely populated with a type of photoreceptive cells, called cones, that are sensitive to small objects, color, and contrast [55]. However, the density of these cells decreases rapidly in the periphery, reducing acuity. The periphery on the other hand mostly contain another type of cells called rods, these are sensitive to light, shade, and motion [55], [62]. The peripheral vision is, instead of providing high-acuity, giving us information [62] about where to look next and what changes or movements that occur in the visual field.

There are three types of positional eye movements, fixation, saccades, and smooth pursuits, that are of interest when observing the visual attention [55]. These movements are defined as follows:

- **Fixations** are tiny movements resembling random noise no larger than 5° visual angle that are stabilizing [55] over a specific area of interest and are said to correspond to one's desire to maintain the gaze on a specific object. These movements range between 150–600ms in duration and about 90% of the viewing time is spent on them [55].
- **Saccades** are, according to [55], rapid eye movements, ranging between 10-100ms in duration, that are used to reposition the fovea to a new location such that a new area of the environment can be visualized. These movements occur as both corrective adjustments of the eye as well as voluntarily controlled eye movements [55] that are used to change the focus of attention.
- **Smooth pursuits** are movements that are used to visually track a moving target [55] and refers to the fact that the eyes, depending on the target movement range, are able to keep up with the velocity of the target.

Other, nonpositional, eye movements are adaptation and accommodation [55] (i.e., pupil dilation, lens focusing).

Detection of Fixations and Saccades

The main goal of ET is, according to [55], to distinguish between the three positional movement types mentioned above. This is done through the localization of regions where the ET signal switches between two stationary values, fixations, where the sharp edges of the changes are the saccades. There are several metrics that can be used to extract further information from the fixations and saccades, e.g. [55] fixation duration, fixation

count, saccade amplitude, and saccade count.

There are mainly two automatic ways [55] to perform this analysis, the first one being averaged summations and the second one is through differentiation. The first one, also referred to as the "dwell-time" method, averages the ET signal over time and if it remains within what can be seen as low variance for longer duration than a specific threshold it is classified as a fixation [55]. The second method, on the other hand, subtracts consecutive data points to estimate the velocity of the eye movements [55], which requires that the ET is performed using a fixed sampling rate. Fixations are extracted from these velocities either as the segments that occur between saccades, or as the segments where the velocity falls below a predefined threshold [55]. There are indications that the second method is better for real-time detection of saccades [55] due to faster calculations. The thresholds, for both methods, are often determined through empirical studies [55].

One of the main issues of ET analysis it that the recorded signal is inherently noisy [55] due to the eye's constant movements and also as a result of eye blinks. Filtering the data before it is used is therefore of importance. Eye blinks should, according to [55], generally be easy to distinguish since they create a large disturbance in most eye trackers.

I-VT filter

The velocity-threshold identification (I-VT) filter is a spatial (velocity-based) algorithm [63] that is used to distinguish fixations from saccades in eye gaze data. The intuition behind the algorithm is that fixations have low velocities ($< 100°$/sec) while saccades have high velocities ($> 300°$/sec) [63]. The I-VT algorithm works as follows [63];

1. Calculate point-to-point velocities.
2. Classify each point as either a fixation, if its below a specified threshold, or as a saccade if its above it.
3. Group consecutive fixations together and calculate the center point of each group based on the center of mass.
4. Set the start time for the fixation as the time of the first point in the group and the duration of the fixation as the time between the first and last point in the group.

The velocity threshold that is used is the only parameter that needs to be specified [63] and it can be set to what is considered a reasonable angular velocity based on computation of angular velocities (requires the distance from eye to visual stimuli to be known) or simply using the sampling frequency in conjunction with empirical data. A study by [64] shows that a threshold between $20 - 40°$/s are suitable values to try for specific eye trackers whereas a threshold of $30°$/s may be a suitable trade-off when working with a multitude of eye trackers.

Other things to consider for the use of the I-VT filter, apart from the selection of the threshold, are [65]:

- **Lack of smooth pursuit detection** - There is no distinction between fixations, or saccades, and smooth pursuits [65] and the latter will therefore always be classified as either a saccade or a fixation, depending on the velocity threshold.
- **Noisy data requires filtering** - All systems that are designed to perform measurements are generally noisy to some extent [65], these disturbances may come from the

equipment as well as from the environment. The way the eye movement velocities are calculated in the I-VT filter, as the fraction between the difference in angular position and the sampling frequency [65], means that if the eye tracker makes even the smallest miscalculations this will introduce significant noise in the velocities calculated from data collected at a high frequency. On the other hand, with eye trackers sampling at lower frequency, the noise introduced by measurement issues will typically still have the same amplitude as for higher frequencies, but as the time between each sample is greater, applying a filter introduces the risk of distorting [65] the original gaze data. Noise generally appears like random spikes in the data [65] and since it has a higher frequency than the signal the I-VT filter aims to detect, it is possible to reduce the noise using a low pass filter [65] that smooths the data by removing signals of high frequency. An alternative to the low pass filter is to calculate the average eye movement velocity over several samples, which is less sensitive to noise than using just two measurements [65].

- **Gap fill-in** - Another issue is that some loss of data is almost always present in digital measurement systems [65] occurring when a sample cannot be collected as the measurement is performed. When it comes to ET in a worn eye tracker this is mostly caused by the participant blinking, resulting in gaps of a few hundred milliseconds. Other reasons that gaps appear, for shorter durations, include delays in data transfers, temporary reflections caused by prescription glasses [65], etc. This could potentially split a fixation in two [65] if the data is not replaced by valid information and one does, therefore, need an algorithm that fills in the gaps.
- **Eye selection** - The eyes are often behaving slightly different when it comes to the start and end time of fixations and eye blinks [65]. This may lead to gaps in the data from one of the eyes and this requires a decision to be made regarding how the data from both eyes should be merged into a single data set for the I-VT filter [65]. Two examples are; averaging between eyes or using only the left or the right eye as the base for calculating fixations.
- **Close fixations** - Imperfections, such as short gaps or noise, results in data points being misclassified [65], which in most cases means that long fixation gets separated into two shorter ones with a saccade in between them that is short in both travelled distance and duration. This can be countered thorough a post-processing procedure that merges fixations [65] that are close in time and space.
- **Short fixations** - The basic I-VT filter does not limit how short a fixation can be [65], but due to the cognitive processes that processes the visual information, that occurs during fixations, there is a limit to how short these can be. This requires the implementation of a filter that removes data points [65], labeled as fixations, that last too short time.

CHAPTER 5

Supervised Machine Learning

The use of modern technology, such as sensors and computer programs, makes it possible to collect more data at a higher accuracy and a higher pace than previously. It is however difficult to analyze these large datasets, sometimes referred to as Big Data [22], using traditional methods.

Machine learning (ML), is on the other hand, a tool that can be used to process these huge datasets and solve practical problems using statistics and probability theory [66]. Supervised ML and unsupervised ML algorithms are the two most common types of algorithms [67]. The former means that the algorithm learns from examples of the output that is expected from a given input, i.e. the algorithm is given labels or targets for each input [67], whereas the latter type of algorithm lacks this information, for example in the task of clustering, where the goal is to retrieve information on underlying patterns or to group data into categories [68].

An ML algorithm generally consists of the following components; a model, a cost function, and an optimization algorithm [67]. These are then coupled with a dataset to solve a specific problem. Different types of learning tasks are, for example, classification, regression, machine translation, anomaly detection, and denoising.

The main challenge in the field of ML, according to [67], is to train a model that performs well on previously unseen inputs, which is called generalization, and not just on the samples that were used during training. During training of an ML model there is a training dataset. This dataset is used to determine the model's performance, called training error, and the parameters of the model are then altered in order to reduce the error [67]. This can be seen as an optimization problem, and what separates ML from optimization is that the goal is to also obtain a small generalization error [67], i.e. the estimated performance on the test

set, a dataset collected separately from the data used in training. The performance of an ML model is, therefore, dependent on the model providing a small training error while it at the same time keeps a small difference between training and test error. The two factors corresponds to two important challenges in ML [67]: the first one, underfitting, occurs when the model is not able to achieve a low enough training error and the second one, overfitting, occurs when the discrepancy between the training and test error is too large. The likelihood of a model to under- or overfit can be managed through the alteration of its capacity [67], which can be seen as its ability to fit a large variety of functions, for example by adjusting the model's width and/or depth. Models with a low capacity may experience difficulties learning the training set whereas the ones with high capacity memorizes properties of the training set, that are not transferable to the test set.

The following sections in this chapter will cover some ML approaches in the realm of ANNs and some of their areas of application.

5.1 Feedforward Neural Networks

Feedforward neural networks (FNNs) are the basic building blocks for deep learning models [67]. The goal of an FNN is, as described in [67], to approximate some function f^*, for example, $y = f^*(x)$ that maps an input x to a category y. An FNN defines a mapping $y = f(x; \theta)$ [67] and learns the value of the parameters θ that gives the best approximation of f^*. Feedforward comes from the fact that the information in these models only flow in one direction [67], from the input x, through the intermediary computations that defines f, and finally to the output y. No information from the outputs [67] is fed back into the model again, when FNNs are extended to include feedback connections, they are called recurrent neural networks (RNNs), further described in Section 5.3.

The "network" component in FNN comes from the models typically being composed of many different functions. One example, given by [67], is the chain of the three functions $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$, that forms $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. This is a common structure in NNs [67] where $f^{(1)}$, in this case, is called the first layer, $f^{(2)}$ is called the second layer, and so on. The total length of the chain is what determines the depth of the model and this is what inspired the "deep" part of deep learning [67].

During the training of NNs, the goal is to make $f(x)$ match $f^*(x)$ as closely as possible [67] using the training data that provides noisy, approximate examples of $f^*(x)$ evaluated at different training points. Each data point, x, has a corresponding label $y \approx f^*(x)$ [67] and the model shall, for each x, produce a value from the output layer that is close to y. The algorithm must learn to decide how to combine the intermediary layers, and the output layer, to approximate $f^*(x)$ as good as possible [67]. However, the training data does not contain any information regarding the desired output from the intermediary layer, which is the reason that they are called hidden layers [67], and the dimensions of these hidden layers determines the width of the model. Each layer can be seen as composed out of several units, acting in parallel, each representing a function that transforms a vector to a scalar [67]. The units are similar to neurons in the way that they take inputs from multiple other units and uses that to compute their own activation value [67]. This, and the fact that the networks contain features loosely inspired by neuroscience, is why they are called neural. However, FNNs should be seen as ways to approximate functions rather than as models of how the

human brain operates [67].

5.2 Convolutional Neural Networks

CNN:s are a type of feedforward neural networks that are more robust to shift, scale, and distortion invariance [69] than fully connected neural networks, and therefore better at detecting spatial and temporal features. This is achieved by convolving or sub-sampling the input to the layer with local receptive fields [69] (filters) of a given size [n x m]. Each filter has $n \cdot m$ number of trainable weights + a trainable bias and these are shared [69] for all filter outputs.

5.3 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a subgroup of ANNs that are used to process sequences of data [67]. An RNN shares its weights across several timesteps [67] wheras a fully connected neural network would have separate weights for each part of a sequence. In an RNN, the current timestep is not only computed as a function of its input, which is the case for regular feedforward neural networks, but also uses previous output states [67] that gives the network access to historical data and how these change over time. RNNs generally also allows for processing of sequences of variable length.

Long-Short Term Memory

Traditional RNN:s tend to suffer from problems with exploding or vanishing error gradients [67], [70] that prohibits proper learning over longer time instances. Long Short-Term Memory (LSTM) cells [70] are designed to provide a solution to this problem using a constant error flow [70] through the network, together with three gates that open and close in order to access the error flow [70]. The input gate decides when the internal state of the LSTM cell should be affected by the input to the cell, the forget gate determines when the cell's internal memory should be reset, and the output gate controls whether the current state of the cell should influence the error flow or not [70]. An LSTM network may contain multiple cells and the network learns to control each individual gate [70] in each cell. The GRU-unit [71] is another type of gated cell that is similar to the LSTM, however, it uses only two gates, a reset gate that determines when to ignore the previous state and an update gate that decides if the state shall be updated or not.

5.4 Dropout

Dropout is a deep machine learning method that is used to reduce overfitting [72]. This is done by randomly ignoring, with probability p, each neuron in a network every time a training case is presented to the network. The goal of randomly excluding some neurons for every training case is to make sure that the network learns generalized features instead of a co-adaptation between neurons [72]. The probability to be used for fully connected layers, suggested by [72], is $p = 0.5$.

Dropout as a Bayesian Approximation

The dropout method described above can, according to [73], be used to approximate Bayesian inference. This is done by enabling dropout at all times, not only during the training of the network, which means that the network will randomly omit some neurons also when making predictions causing variation. The mean prediction as well as the model uncertainty can be obtained by making N number of predictions [73] on the same data and collect the results. [73] claims that $N \in [10, 1000]$ should give reasonable results. Using this approach is useful since it provides a way to reason about model uncertainty that is easy to implement and less computationally expensive [73] than alternative methods. [73] suggests that the probability p for dropping a neuron should be in the range of $p \in [0.1, 0.5]$.

5.5 Alternative Neural Network Architectures

There exists a wide range of other network architectures that can be used to analyze sequences of data. A few examples are; Transformers [74] that have been used for translation tasks and is based on a mechanism that learns what inputs that should be given attention at each point in time, auto-regressive neural networks such as PixelCNN [75] and Wavenet [76] that has been used to generate images pixel-by-pixel or raw audio respectively, and graph CNNs, e.g. STGCN [77], that has been used in human action recognition tasks that model the human body using skeleton key positions.

CHAPTER 6

Experimental Procedure

This chapter defines an experimental procedure that can be used to analyze human behaviour using VR, ET, and ML. It starts off with the formulation of the objective, what data is needed, and how the performance of the ML solution will be evaluated. This is then followed by three steps that describes how VR with ET can be used to collect human movement data, namely the experimental setup that covers the hardware, the test development that describe how a VRE can be designed in order to collect gaze and movement data, and then the test study and the selection of participants will be covered. The test related sections are followed by suggestions for ways to preprocess the data before use with a ANN, the design of a ANN architecture that may solve the objective, and what to consider when the results are obtained. Each section will also provide a brief description of what has been used in **Paper A-D**.

6.1 Objective - What is of interest and why?

The first thing to consider is what problem(s) is(are) of interest, how these may be solved through a ML approach, and what is the end goal or the final product. Once the objective is clearly defined it might be useful to identify possible subgoals that can be used to explore the overarching objective through an iterative process that may provide partial solutions. Both the main objective and its subgoals need to be measurable in such a way that they can be evaluated in a meaningful way. Evaluation will be further explored in Section 6.3.

The objective in **Paper A** was to investigate whether human eye gaze data can be used to classify which object out of 5 boxes that was selected after the test procedure was

completed. This was slightly modified for **Paper B** where the aim was to try to predict at what angular direction the test participant was going to position their hand a fixed step ahead of time (500ms). The objectives from **Paper A** and **Paper B** are two subgoals that together provide the foundation for **Paper C**, which had the objective to continuously classify, ahead of time, which box, out of 18 possible ones, that the test participant is about to select based on the ANN's uncertainty.

In the extension to **Paper D** the goal was once again to classify what object was selected, but in this case what alternative that the user selected as the correct alternative for a logical pattern that does not require hand movements.

6.2 Data - What data is needed?

Once the objective is clearly specified it is time to figure out what data is needed to build the ML model in order to provide a solution to the problem. This involves the following steps:

- If the data is not already available, it has to be gathered somehow and this is both a time consuming and possibly costly procedure if it involves, for example, human test participants.
- The amount of data required to develop a working ML solution varies and one must determine the minimum number of data points that gives a working solution. However, the quality of the data is also of importance and this requires good equipment, a suitable test design that accurately represents the objective that is to be solved, and that the data points with the most valuable information are collected. In order to make the most of the data during the training procedure one might also want to employ some type of data augmentation, further described in Section 6.7.
- The next step is to figure out what measurements are of interest, how and to what degree these capture the different aspects of the problem, and how these may be used to solve the objective.
- The available architectures and the possible problem formulations are also affected by the type of data that is collected and how it is arranged, e.g. sequences of numerical data, matrices, images, etc. The formatting of data will be covered in Section 6.7.
- The problem formulation will also affect the required data, for example a many-to-many sequence problem might require as many labels as input data, whereas a many-to-one classification problem might be solved with a single label for each set of data points.

The data that is used in the four studies in Chapter 7 has been collected from volunteer test participants and in most cases with the author of the thesis as the test leader. The dataset in **Paper A** features 720 data points collected from 24 participants, **Paper B** has a dataset of 8512 data points that was obtained from 14 participants, and the dataset used in **Paper C** contains 3192 data points from 21 participants. The extension to **Paper D** uses a dataset that was collected by [78] and contains 1840 data points, obtained from 184 test participants. These datasets can be seen as quite small in ML context, however, the results shows that small datasets might be sufficient to develop proof of concept, especially for tasks of lower complexity.

The type of data is almost the same for all studies. It involves eye gaze, HMD movements, controller movements, and test specific data, such as selected boxes or alternatives, which is were the differences come into play. **Paper A**, **Paper B**, and **Paper D** employ a many-to-one problem formulation whereas **Paper C** is trained as a many-to-many problem that is used in a many-to-one context.

6.3 Evaluation - How will the network results be evaluated?

In order to accurately determine whether the ML solution successfully solved the objective or not it is important to know beforehand how to evaluate the networks performance. Depending on the problem to be solved there may be different ways of doing the analysis, for example if it is a commonly or previously tried experiment one should consider using the same metrics or benchmarks in order to make it possible to quickly do comparisons. In other applications the use of common metrics such as mean squared error or classification accuracy etc., may falsely evaluate the network due to a system dynamic that is not clearly captured in the metric. An example of this could be to incorporate slack in the error measurement if the system itself does not require pinpoint precision.

Paper A, **Paper C**, as well as the extension to **Paper D** used a traditional classification accuracy score in conjunction with prediction filtering based on the standard deviations of the predictions that allows the system to be evaluated according to its certainty as well as performance. **Paper C** was evaluated slightly different to the other two in order to simulate the performance on a continuous stream of data. **Paper B**, on the other hand, used a solely custom metric that gives the system some slack since the application only requires the estimates to be in an angular area rather than at a specific angular value.

6.4 Experimental Setup - What hardware can be used to collect the data?

There several ways of measuring eye gaze, e.g. camera-based ET methods [79] and wearable eye-tracking glasses [80], as well as human movements, e.g. camera-based methods [81] and wearable inertia based methods [82]. One way to merge gaze and movement tracking into one system is through the use of a VR-headset with hand controllers and built-in ET. It is possible to design a fully controllable VR environment (VRE) that gives access to information about where the user has been looking, moving their head and hands, while simultaneously limiting visual distraction through the immersion that the headset gives. Performing various experiments in VR is also much safer than in the real world since there is no risk that the operator is hurt nor that the equipment, inside the VRE, is damaged. There is also an endless supply of material since generating new parts is simply a piece of code.

Paper A-D all use the same setup, namely a consumer grade VR-headset, "*Tobii Eye Tracking VR Devkit*" [83], that has built-in ET and utilizes two handheld controllers to navigate the VRE.

6.5 Test Development - How can the test be designed with regards to the data objective and available hardware?

Developing a VRE test can be broken down into the following steps:

- Language - Choose one or multiple language options that can be used to present written instructions during the test procedure and make sure that the test leader is able to deliver the spoken instructions in the chosen languages. The instructions should preferably be customized to fit the target group that is going to perform the test.
- Design the test in a way that makes it as clear as possible to follow the different steps of the tests including for example; language selection, calibration of the equipment, input of extra information such as age, gender, etc. (to use for basic demographics), and the start of the test itself. It is also convenient if the test is easy to restart if something goes wrong, the participant has additional questions during the test, or simply to make it easy to move on to the next participant.
- Consider using anonymous participant IDs in order to store the collected data in a way that preserves the privacy of the participants.
- Limit distraction - Limiting or controlling distractions from the test itself is good way to reduce or introduce noise in the data depending on what is desired. Visual stimuli is easily controlled in a VRE and this is an important strength that should be utilized to make the test as standardized as possible.
- Warm-up - If the test procedure is unfamiliar to the participant then it might be useful, in order to reduce the bias from inexperience with for example the equipment, to have a warm-up segment that ensures that the participant gets some experience regarding what is to be done.
- Finally, consider developing the VRE in a modular fashion such that modifications can be made if necessary or in order to be able to reuse the parts of the environment in a different context.

The main VRE structure for **Paper A-D** uses Swedish and English as the two instructional languages, it incorporates anonymous participant ID:s that are randomly generated as the test is launched, the ET calibration steps and the gathering of general information follows the same procedure, and its ET, movement tracking, and visual distraction limiting features are the same. This makes for a modular design that is slightly modified as the objective changes through the papers.

The test stage in **Paper A** features a table in the form of half a circle where cubes will appear at random in 5 different zones with 45° spacing. The test stage has been designed in a way that is meant to force the test participant to look in the direction of the cube, make a movement towards the cube, and acknowledge that movement by touching the cube. The zone that gets a cube is randomly selected every time a new cube is to be created and the positioning of the cube is also randomized, in the interval $x \in [-x_s, x_s], y \in [-y_s, y_s]$, where x_s, y_s are the maximum deviations allowed around the center of each zone. The test

has a 1s time delay between each cube appearing that helps to slow down the pace of the test execution.

Paper B and **Paper C** use the same test environment featuring two even distributions of 9 cubes, each at two different heights and radii that allows for a flexible design of a sequence of movements; forward, backward, left, right, sweep left, and sweep right. The cubes appear at two different radii, based on the participant's arm length, and requires the test person to touch it while simultaneously pressing a button on the controller to make the cube disappear. After a cube has disappeared, and a delay of 0.2s, the next cube in the pre-defined sequence is lit. The delay is used as a way to force a slower pace throughout the test and data is collected during this time. The only differences between the tests are that the movement sequences in **Paper C** are randomized, and the total number of movements are less as a result of feedback from participants stating the test as to long and strenuous, and data is collected during the delay.

The VRE in **Paper D** is designed to model different items from Raven's Progressive Matrices and takes place in a sparsely furnished, square space with calm colors to prevent the user from being distracted. The user can move freely in the room throughout the test, both in the real world and in the virtual, but it is recommended to remain seated/standing still.

6.6 Test Study - How will the data collection take place and who will participate?

- Limit external distractions - External distractions or disturbances may have a negative impact on the quality of the data since that could give some of the participants unfair disadvantages. It is therefore crucial to limit these (the ones that have not already been taken care of during the test development) as much as possible, for example by using a room with low noise levels or making sure that the participants does not feel stressed about the upcoming task. The latter may be reduced by clearly explaining the goal of the test, going through what is expected of them, and answer any questions.
- Standardized instructions - Make sure to use the same instructions for every participant in order to reduce bias from the instructional phase, however, keep in mind that some people may need some additional help in order to be able to carry out the test as intended.
- Feedback - Ask the participants if they are willing to give some feedback that can be used to improve the test procedure and/or the VRE.
- Selection of participants - In order to create a robust system that works in various conditions and for as many users as possible one needs to consider the test group diversity. It could be possible to design a system that successfully learns how a few participants behave, that is not transferable to others. A larger test group mitigates this as well as other biases towards, for example, age or gender. The importance of this may, however, vary depending on whether the goal is to create a system as a proof of concept or if it is supposed to be production ready.

The data in **Paper A-D** has been gathered at mostly quite places, however, there have been occasions where there have been other people present . The instructions, for each test study, have been given in the same way to every participant. The test procedures and the VREs have successfully been improved through feedback, from early participants, before the actual test studies were carried out and between the studies. **Paper A-D** are all simplified test studies that are designed as a proof of concept. The selection of participants is, therefore, limited to volunteers who work or study at Chalmers.

6.7 Preprocessing - How will the data be processed to fit the ML solution?

In order to train the ANN with the most useful information possible it is important to preprocess the data such that it is presented in its most usable format. This includes selecting the appropriate labels to be used for both training and evaluation to determine the success of the network. The following steps should be considered:

- Filtering of outliers - If the data consists of unwanted outliers these should be removed before training in order to reduce the likelihood that these guide the training of the network in the wrong direction.
- Selection of features - The features are the input to the ANN, and selecting these will greatly influence the success of the training of the network. Features that are too similar may for example drive the training of the network into a local optimum due to an over-representation of redundant information. Analyzing the correlation between different features could be one way of determining which ones that provide valuable information.
- Data augmentation - DL generally requires a lot of data and if it is difficult and/or expensive/time consuming to gather more data, for example when dealing with human test participants it may be possible to augment the training data in order to achieve a better performance. Ways of augmentation could be to add duplicates of the data, with or without noise, shuffling of the data, and other transformations that slightly perturbs the data such that it aids the networks generalization capabilities.
- Network specific preprocessing - Depending on the task and the network architecture the data may need to be formatted in a specific way in order to solve the objective. If sequences require equal length one may consider for example zero-padding the data or applying some kind of up- or downsampling to give them equal length. Other contexts could require that the data is parsed using a sliding window.
- Normalization/standardization - The network may over emphasize the importance of some features over others if they are of different magnitude. This can be countered through feature wise normalization or standardization of the data. The former referring to re-scaling the data, for example between its minimum and maximum value, and the latter to re-scaling the data to have zero mean and unit variance.
- Simplifying the problem - Before it is time to select the labels for the supervised learning problem one may consider reformulating the problem to reduce the number

of different variations that the network has to learn. One example of this in a classification problem could be if there are 2 classes that contain 10 similar subclasses. This could either be formulated as a multi-class problem with 2*10=20 classes or as a binary classification problem coupled with a multi-class problem with 10 classes.

- Selecting labels - Supervised ML requires suitable labels to learn how to solve the objective. Selecting the appropriate labels is the difference between a successful and a failed project. A clearly defined objective and evaluation procedure should, therefore, be the starting point for the selection of labels along with a careful analysis of the available data.
- Class re-balancing - The data that is collected may sometimes be unbalanced, i.e. there are one, or a few, classes that constitute the majority of the data. This may cause the network to become biased towards guessing these classes and in worst case rendering the network useless since it learns to always guess the majority class, as the cost of a wrong guess is small compared to all the right answers. Class unbalances can be mitigated through re-weighting the loss of being wrong during training, such that the loss of guessing wrong is higher for rare classes. It could also be possible to collect more data that is targeted towards the minority classes or, if it is possible, redesign the test procedure to ensure balanced data.
- Train/validation/test split - The last step of the data preprocessing is to split the data between training and test data. The size of these may vary but the important thing is that the test set is large enough that it is possible to determine that the network generalizes well, while on the other hand a larger training set usually improves the training procedure and reduces the risk of overfitting, thereby providing a better generalization. A common strategy when experimenting with network parameters is to also split the data a third time into a validation set. This is useful both during the training phase to monitor the loss on unseen data, but also to mitigate the risk that repeated experiments make the solution tailor-made towards the test set, which should only be used for the final evaluation.

Paper A-D use the same way of discovering and removing outliers. The datasets have been approximated using Beta-distributions and then a maximum threshold, maximum duration (samples) of a test segment, has been set according to the mean plus three standard deviations of this distribution. All data points that contained more samples than the threshold were discarded.

The next step filters out each sample, within each data point, that contained NaN values. These data points were discarded in **Paper A** whereas they were replaced with the previous valid sample for **Paper B-D**. NaN values occur when the ET fails to read the eye properly, most common as a result of the participant blinking.

In **Paper A**, one of the goals was to investigate the neural network's ability to handle raw gaze data and the features that were used are, therefore, left and right gaze direction vector, left and right pupil diameter, and a variable that contains the duration of the test. The same features, apart from the exclusion of the pupil diameter, were then used in **Paper B** as well. **Paper C** analyzes the correlation between the features that were available and based on this uses the average gaze direction vector, two of the HMD position coordinates, and the average pupil diameter. In the extension to **Paper D**, the average gaze direction vector is used together with a boolean variable that tells the network whether there are 6

or 8 alternatives available for the user to select.

Paper C is the only study where data augmentation was added and it is achieved by stacking randomized movements after each other in the creation of the training dataset, thereby creating new transitions between movements since the ordering is different.

The data points in **Paper A-D** that were of shorter length than the decided threshold were padded with zeros (ZP) at the end to guarantee the structure of the data point that is fed to the network. **Paper A** also evaluates using linear upsampling (US) of the data points to achieve the desired length. US is, however, not an alternative for a continuous data and since ZP performed better, it was chosen as the preferred method. After ZP, the data was normalized featurewise between -1 & 1 or using the max-norm.

The problem formulation for **Paper A** and **Paper D** is simply a multi-class classification problem that uses either the id of the boxes or the id of the alternatives as labels. **Paper B** is formulated as a regression problem where each label is an angular value. **Paper C** was rewritten from a 19 class classification problem to a 10 class problem, with the ids of the boxes at the lower level are the labels. Nine of these classes, the box in the centre is excluded, are also binary classified as either 0 or 1, corresponding to the lower or the upper level of boxes.

Neither of the papers use any class re-balancing and all training procedures, in **Paper A-D**, used roughly a 45%/5%/50% (train/validation/test) split.

6.8 Neural Network Design - What network architecture(s) can be used to solve the task?

Once the data is formatted in the proper way it is time to create the ANN that will perform the analysis of the data. A first step is to consider if there are any specific ANN properties that are suitable for the specific task, such as CNNs for images or fixed time series, or RNNs for more complex time series problems. Start off with a simple network and add more complex structures later, this makes the network easier to analyze if it is not working. It trains faster, and lowers the computational cost. This is also when one may consider whether there are any hardware limitations that come into play when using the trained network in its live environment. Fast online predictions are easier to achieve using a smaller network since it requires less computational resources. The choice of intermediary activation functions may affect the performance of the network [67] and a few common alternatives are; `ReLU` [84], `tanh`, and `sigmoid`.

When the network architecture is in a place one needs to apply an output activation that transforms the network output to its desired format [67] and then couple it with the appropriate loss function that controls the learning process of the network. A `softmax` output is commonly used for multi-class classification and is often paired with a `categorical crossentropy` loss whereas a `sigmoid` activation is paired with `binary crossentropy` loss for binary classification, etc.

The training of the network also requires an optimizer that is used to find the appropriate error gradients to learn from [67] and some optimizers may work better for a specific architecture than others. It is also important to determine when the network should be considered fully trained. One approach is to monitor the networks performance on the vali-

dation data and terminate the training once the validation loss stops decreasing, sometimes referred to as early stopping. The reason that the performance on the validation data is considered is because the objective is to train a network that works well with unseen data and not to optimize on only known data.

Paper A and the extension to **Paper D** uses a CNN approach, inspired by the inception modules from `Inception-v3` [85], adapted to 1D time-series data as the basis for the classification coupled with the uncertainty estimation described by [73]. The architecture, Figure 6.1, is utilizing `ReLU` activation functions, a `softmax` output activation, and is trained with `categorical crossentropy` loss. This structure is then reused for **Paper C** in a time distributed way with the addition of an LSTM-layer, Figure 6.3, and the intermediary activations have been swapped from `ReLU` to `tanh`. The network has two outputs, the first one uses a softmax activation together with a `categorical crossentropy` loss whereas the second one uses a `sigmoid` activation paired with a `binary crossentropy` loss. **Paper B**, Figure 6.2, is also utilizing convolutions for 1D data, however, in a less complex architecture that is used to explore the effects of varying depth and parameter counts. The output from this network is real-valued and is therefore using a `linear` output activation, the other activations are using `tanh`, and the network is trained with `mean absolute error` loss. **Paper A-D** were all trained with the `adam` optimizer [86] with default settings until the validation loss stopped decreasing.

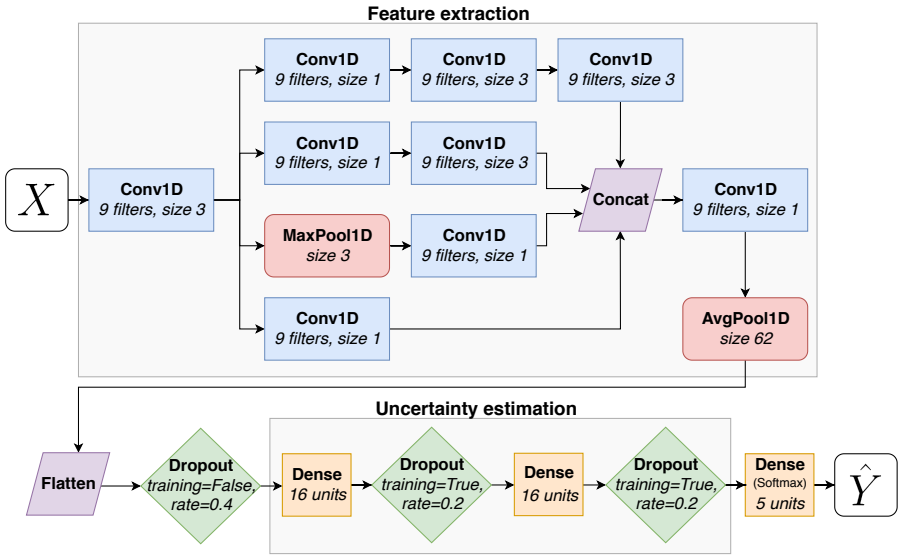


Figure 6.1: A flowchart that describes the network architecture used in **Paper A**.

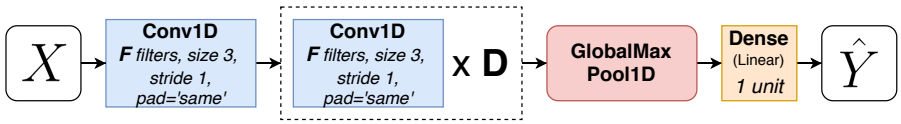


Figure 6.2: A flowchart that describes the network architecture used in **Paper B**.

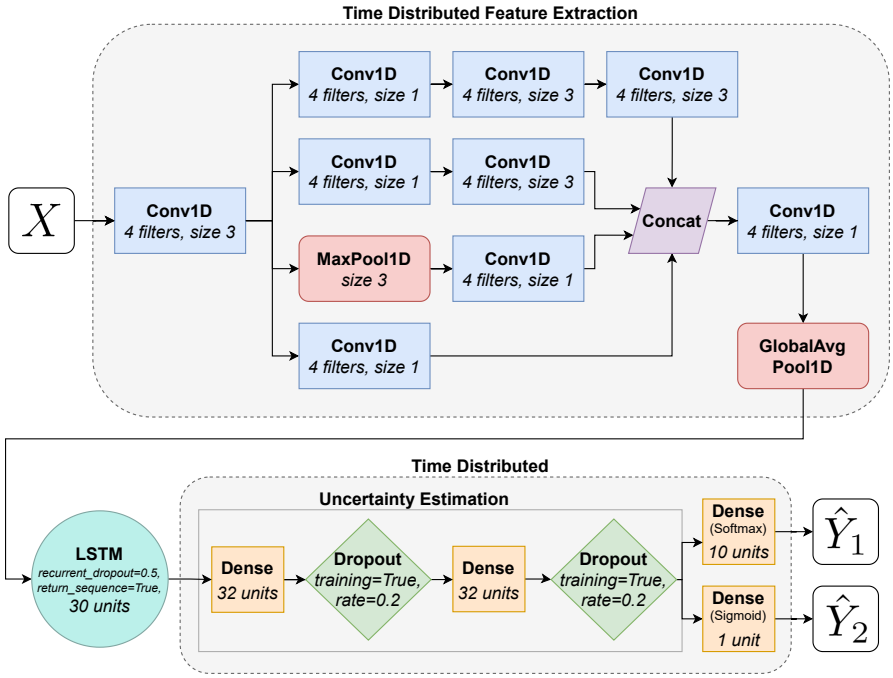


Figure 6.3: A flowchart that describes the network architecture used in **Paper C**.

CHAPTER 7

Human Intention Prediction

This chapter will present the results that were obtained from the three test studies explained in **Paper A-C** that investigates the ability to determine human movement intention, in three stages of complexity, based on eye gaze. These three studies are accompanied by an extension of **Paper D** that utilizes the same tools to perform intention analysis in a different context, namely in the realm of psychological testing where the task is to classify which answer, to a logical pattern test, that the participant selected. The main takeaways include that it is possible to analyze human intent in similar ways regardless of application as long as the eye gaze is used as the input data. It is also clearly visible that modelling the uncertainty of the ANN is greatly improving the analysis and discussion of the networks performance, both from a safety and a usability perspective.

7.1 Study 1 - Human Movement Direction Classification using Virtual Reality and Eye Tracking

This section will provide the classification results, on the test set, from the trained network and a comparison that shows the impact of using UE [73].

The classification results without UE can be seen in Fig. 7.1. The graph shows the largest contributor from the softmax output for each sample that was classified. The samples are sorted in increasing order, left to right, based on this value. A green bar represents a correctly classified sample whereas a red bar indicates that the sample was incorrectly classified.

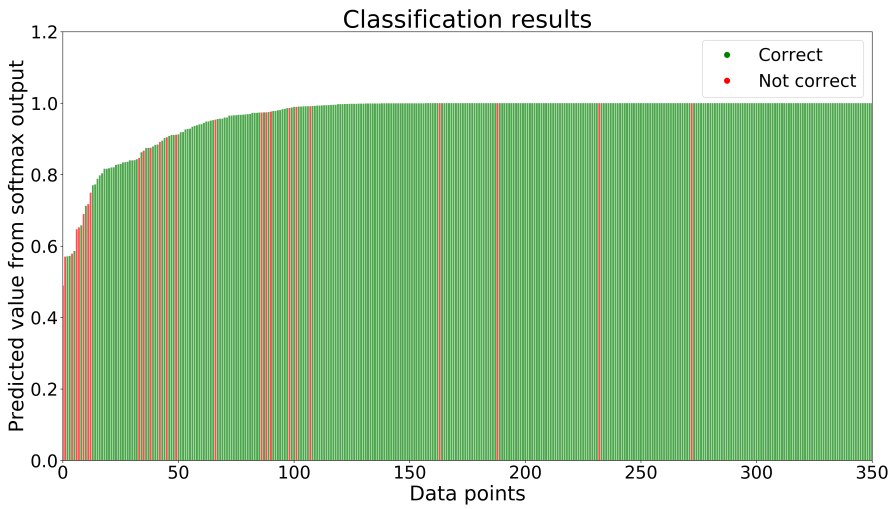


Figure 7.1: A graph of the classification results without UE.

The difference when making predictions for UE is that many predictions are done on the same data such that it is possible to obtain a mean value and a standard deviation of the prediction. The pseudo code for this is shown in Algorithm 1.

The results obtained from the network using UE, with $nrOfPredictions = 1000$ as suggested by [73] to ensure good plots, can be seen in Fig. 7.2. The graph shows the largest contributor from the softmax output for each sample that was classified. The samples are sorted in increasing order, left to right, based on this value. The black interval displays two standard deviations of the prediction around its mean value. A green bar represents a correctly classified sample whereas a red bar indicates that the sample was incorrectly classified.

Algorithm 1 Pseudo code for predicting with UE.

Input: X, nrOfPredictions
Output: $\hat{Y}$, $\hat{Y}_{STD}$

```
predictions = []
for i = 0 to nrOfPredictions do
    predictions[i] = model.predict(X)
end for
Ŷ, Ŷ_STD = mean(predictions), std(predictions)
return Ŷ, Ŷ_STD
```

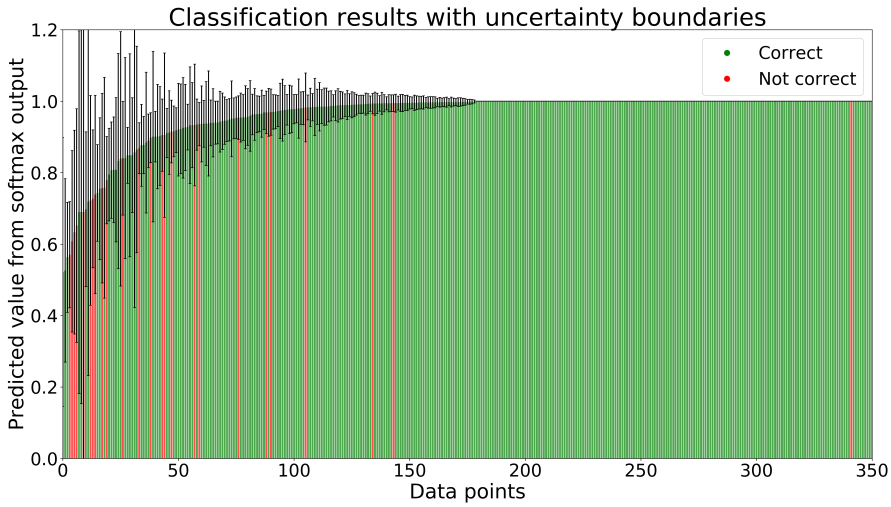


Figure 7.2: A graph of the classification results with UE.

Once the mean and standard deviation has been obtained from the network these can be used to determine if the network is confident enough, high mean and low standard deviation, to make an accurate prediction. This was implemented as shown in Algorithm 2 where a prediction is accepted if the mean minus two standard deviations is larger than a chosen lower limit.

The classification results for different lower limits can be seen in Table 7.1. It is clear that the classification accuracy can be increased with this approach, however, at the cost of the network not being able to classify all samples.

Algorithm 2 Pseudo code that accepts or discards a prediction.

Input: $\hat{Y}$, $\hat{Y}_{STD}$, lowerLimit
Output: $\hat{Y}$
1: **if** $\hat{Y} - 2 * \hat{Y}_{STD} >$ lowerLimit **then**
2: Accept $\hat{Y}$ as the prediction for this sample.
3: **else**
4: Discard $\hat{Y}$, the network is not confident enough.
5: **end if**

Table 7.1: Comparison of classification results for different levels of filtering using UE for both US and ZP.

Lower Limit	US - accuracy	US - % samples classified	ZP - accuracy	ZP - % samples classified
0	88.37%	100.00%	93.28%	100.00%
0.10	89.03%	98.97%	93.52%	99.74%
0.20	89.36%	97.16%	93.73%	98.97%
0.30	91.71%	93.54%	93.70%	98.45%
0.40	93.68%	89.92%	94.44%	97.67%
0.50	94.67%	87.34%	95.39%	95.35%
0.60	94.82%	84.75%	95.59%	93.80%
0.70	95.82%	80.36%	96.31%	90.96%
0.80	96.22%	75.19%	96.76%	87.60%
0.90	96.76%	63.82%	98.33%	77.26%

7.2 Study 2 - Human Movement Direction Prediction using Virtual Reality and Eye Tracking

The performance of the networks has been evaluated using a custom metric that is more suitable to the task than a standard measure of error. It is defined as the network's hit-rate (HR) inside a cone in front of the test person with an angular spread θ_T that can be varied to change the size of the cone. The HR, Equation (7.1), is calculated as the fraction of how many of the N predictions of $\hat{\theta}$ that were less than $\frac{\theta_T}{2}$ degrees away from the target hand direction, θ.

$$\text{HR} = \frac{\sum_{n=1}^{N} \left(|\hat{\theta}_n - \theta_n| < \frac{\theta_T}{2} \right)}{N}. \tag{7.1}$$

A number of networks have been trained using all combinations of values for $D \in [2, 4, 6, 8, 10]$ and $F \in [2, 4, 6, 8, 16, 32, 64]$, which gives a total number of 35 networks. Each one of these combinations has been trained ten different times and the average resulting HR has been used to evaluate the performance of a combination. This has been done in order to reduce the possibility that a lucky training session made a certain combination successful. The results from the eight parameter combinations that performed the best and the eight worst ones, for θ_T=20°, are shown in Table 7.2. This is the most interesting threshold since the nine cubes in each layer are spread over an arc of 180°, which gives each cube roughly 20° of space. It is clearly seen in Table 7.2 that a parameter count above 10^4 does not improve the HR, however, the deviations in accuracy between all models are small, and the combination that performs the best is F=16 and D=2.

Table 7.2: Table showing an HR-comparison of the 8 best models and the 8 worst models, for several values on θ_T, sorted based on the performance when θ_T=20°.

Model	Params	θ_T=10°	θ_T=20°	θ_T=45°	θ_T=90°
F16-D2	$1.9 \cdot 10^3$	**39.90%**	**62.50%**	86.80%	96.43%
F16-D4	$3.5 \cdot 10^3$	39.53%	62.44%	86.69%	96.45%
F64-D2	$2.6 \cdot 10^4$	38.94%	62.30%	86.78%	96.51%
F32-D2	$6.8 \cdot 10^3$	39.02%	62.29%	86.77%	96.51%
F16-D6	$5.0 \cdot 10^3$	39.23%	62.28%	86.67%	96.41%
F16-D8	$6.6 \cdot 10^3$	39.11%	62.27%	86.61%	96.39%
F32-D6	$1.9 \cdot 10^4$	38.67%	62.21%	**86.87%**	96.52%
F8-D2	$5.6 \cdot 10^2$	39.62%	62.18%	86.59%	96.29%
⋮		⋮		⋮	
F64-D8	$1.0 \cdot 10^5$	37.90%	61.47%	86.61%	**96.56%**
F2-D10	$1.8 \cdot 10^2$	39.64%	61.32%	86.02%	96.13%
F4-D8	$5.0 \cdot 10^2$	39.82%	61.27%	86.07%	96.23%
F64-D10	$1.2 \cdot 10^5$	37.10%	61.25%	86.51%	96.51%
F2-D8	$1.5 \cdot 10^2$	39.67%	61.07%	85.95%	96.15%
F2-D6	$1.3 \cdot 10^2$	39.52%	61.07%	85.95%	96.19%
F2-D4	$9.7 \cdot 10^1$	39.56%	60.95%	86.04%	96.17%
F2-D2	$6.9 \cdot 10^1$	39.33%	60.81%	85.94%	96.11%

7.3 Study 3 - Human Arm Movement Intention Prediction using Eye Tracking

The performance of the networks in this study has been evaluated using the following custom metrics;

- A_P = Accuracy of predictions that are above UE threshold,
- A_M = Accuracy of how many movements are correctly identified at least once,
- M_T = Mean fraction of time left until the completion of the task,
- A_{VP} = Vertical accuracy, evaluated whenever there is a box prediction.

These were considered more suitable to use to evaluate the network on how well it is able to utilize its notion of UE in order to predict the intended movement direction, compared to a standard accuracy metric that does not capture the aspect of UE at all.

The evaluations have been performed in a way that imitates the continuous flow of data in a real world system. This is done by making a prediction for every timestep of the test set, starting with all zeros as the input and then shifting the input data by one at a time in order to "obtain" new information. The last of the N_w predictions at each timestep is the one that is evaluated since that refers to the most current timestep.

The first network configuration that was evaluated, on the validation data, was $w = 350$ and $N_T = 1$, the results can be seen in Fig. 7.3. The validation data set contains 27323 timesteps in total and the performance was evaluated using $n = 25$ for the UE, and obtained using the same UE approach and filtering as described in **Study 1**, Section 7.1. The intersection point between the two metrics, A_P and A_M, can be seen as the networks optimal performance since the accuracy for the predictions are high while at the same time, as many movements are covered as possible. This is, therefore, the point that specifies the threshold, Th_L, that is used to further evaluate the network.

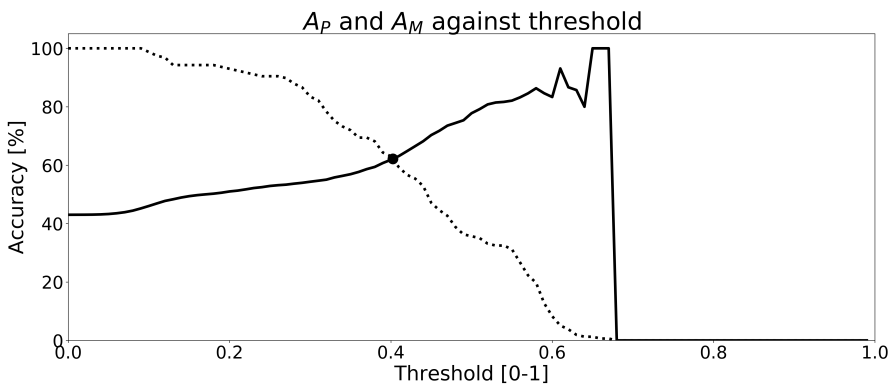


Figure 7.3: A figure showing the selection of the optimal thresholds for each dataset.

Different configurations have then been evaluated in order to determine the affect of changing the subwindow size w_s, the number of training copies N_T, and whether the order of the movements in the augmented data should be randomized or not (NR=No Randomization). The comparison between these are based on the intersection points, described above, seen in Fig. 7.4 where the black dot corresponds to the original configuration. It is clearly seen that changing N_T does not improve the results unless NR is used as well. The best performing combination is $w_s = 10$ and $N_T = 10$ with NR. Different values for w, [175, 700, 1400], was also tested but these showed no significant improvement.

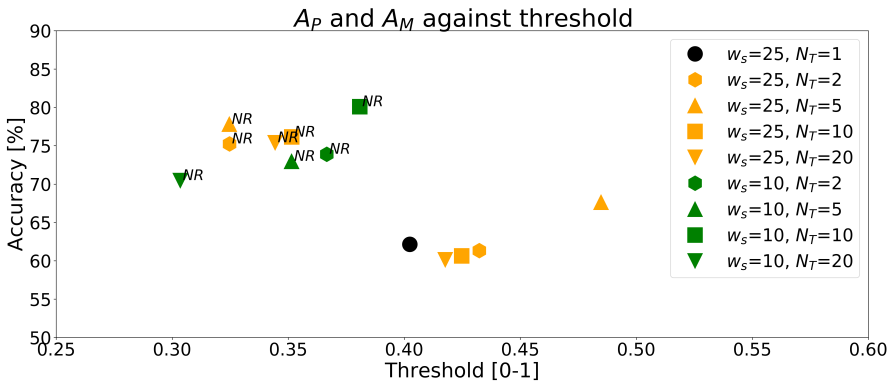


Figure 7.4: A figure showing the comparison of different network configurations, using the intersections of the metrics A_P and A_M. NR stands for not using randomized copies of training data.

The original configuration, $w = 350$ and $N_T = 1$, and the best performing one from Fig. 7.4, $w_s = 10$ and $N_T = 10$ with NR, have been further evaluated on the test set. The test set consists of 250660 timesteps in total and the performance was evaluated using $n = 25$ for the UE. The resulting comparison, using each networks optimal threshold obtained on the validation set, can be seen in Table. 7.3. This table also shows the performance measures obtained using M_T and A_{VP}, the best performing model achieves an average M_T of 37.2% ahead of the action being completed. Using the median length of a movement, this corresponds to 150 samples $* 0.372 \approx 550$ms. The vertical predictions have been evaluated every time a new prediction is accepted as a contribution to A_P, the result for the best model is $A_{VP} = 81.29\%$.

Table 7.3: Table showing a performance comparison between the first network configuration and the best model, evaluated on the test set.

First and best model - Test Set						
w_s	N_T	Th_L	A_P	A_M	M_T	A_{VP}
25	1	0.40	54.18%	50.64%	32.12%	70.38%
10	10	0.38	70.70%	67.89%	37.20%	81.29%

Fig. 7.5 shows a segment of prediction results from, $w_s = 10$ and $N_T = 10$ with NR, obtained on the test. The upper graph displays the predicted and true labels without taking the UE into account. The graph in the middle show what predictions are above the specified threshold and the graph at the bottom show how the certainty, regarding the most probable output class, fluctuates over time.

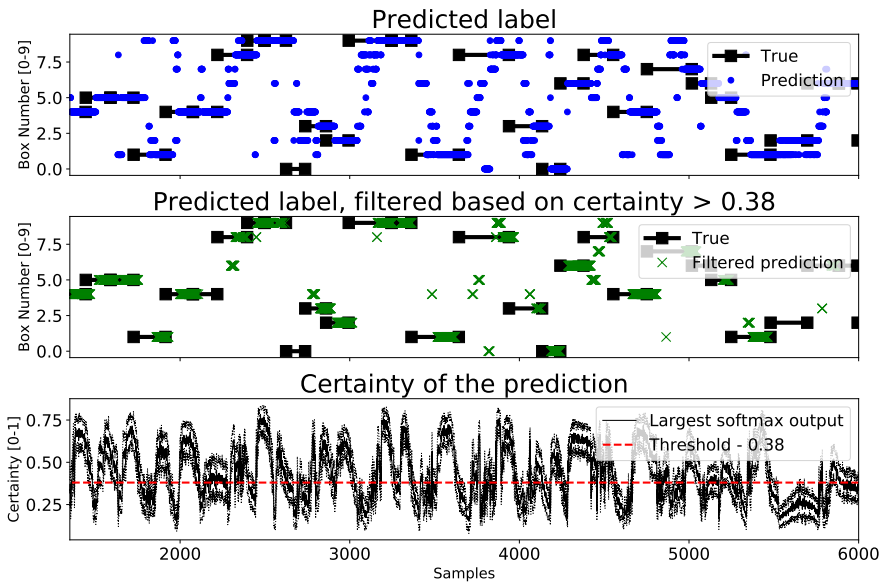


Figure 7.5: A figure showing a prediction segment from the network configuration, $w_s = 10$ and $N_T = 10$ with NR, obtained on the test set.

The computational cost of the model has been estimated to $\frac{1707000 \text{ ms}}{250660 \text{ samples}} = 6.81$ms per UE-prediction (n=25, 0.27ms/prediction) on a 8 core CPU.

7.4 Study 4 - Cognitive Ability Evaluation using Virtual Reality and Eye Tracking

This section will provide an extension to **Paper D** in order to show that the same experimental procedure that was described in Chapter 6 can be applied to a similar, but different, context. The goal of this extension is to determine an objective, identify what part of the data collected using the system in **Paper D** that can be used in an ML solution to solve the objective, and how this can be evaluated. The data will be preprocessed to suit an ANN and suitable features will be specified. This will be followed by a presentation of the developed network architecture from **Paper A**, adjusted to fit the current objective, along with the prediction results that were obtained from the ANN. The steps that describe the experimental setup and test development will not be covered since those are covered in **Paper D**.

1. Objective

The objective is to determine which alternative (1-8) that the test participant chose for each item in RPM that was presented to them, based on eye gaze data alone, using a multi-class classification approach. This should be feasible since the user, according to [78], selects their answer using the gaze.

2. Data

The data that was collected in **Paper D** consists of eye gaze vectors and the alternative that the test participant selected. There are a few other parameters available but, these will not be used since the objective is to classify the selected alternative based on eye gaze.

3. Evaluation

The network will be evaluated using the traditional classification accuracy and the uncertainty filtered classification accuracy that was used in **Paper A**.

6. Test Study

The dataset used in the extension of **Paper D** was collected as part of the master thesis work by [78] and the procedure was described as follows. The data was collected at two demonstration sessions at the Department of Psychology, University of Gothenburg and at one technical fair at Chalmers University of Technology [78]. There has also been some testing of people that were recruited through direct messaging and face-to-face interaction. A week-long test study at the Ågrenska Foundation was also performed. This was carried out by an employee at Ågrenska who was taught to use the entire system without the presence of the authors [78]. The participants at the University of Gothenburg were a mixture of students, teachers and visiting high-school students, whereas there were mainly students participating at Chalmers University of Technology. The participants from Ågrenska Foundation were mainly staff and volunteer workers.

The test procedure at the two universities, described in [78], involved setting up the equipment and the informational poster in an open area where people in general pass through. Information about the research behind [78] and the test was then given to groups of people that stopped by. These areas were often quite crowded with a lot of background noise. The equipment and the poster have also been set up in the authors office during most part of the project. The office is a smaller and quieter space compared to the open areas at the universities. This is where people that have been directly contacted have come to do the test. The environment at Ågrenska Foundation was very similar to the one in the office. The tests were carried out in a smaller, separate room with little disturbances.

All data was collected as follows [78]; the test participant is first of all given a brief explanation of what the work in [78] is about and the purpose of the data collection. After that there are a series of steps that the test instructor walks through to aid the participant through the test. These are described below [78]:

1. The participant is told to put on the HMD with their eyes centered in the middle of the lenses and adjust the fit with the screw on the back of the headset.
2. The instructor hands over the hand controllers.
3. The participant is asked to select the most suitable language, either Swedish or English, using the laser pointer and the touchpad on the controller.
4. The participant is told to stand still and just use eye gaze to complete the calibration step.
5. The instructor informs that the selection of objects in the VRE is now made using eye gaze, but the final choice is still acknowledged using the touchpad on the controller.
6. The participant is instructed to complete the information form that is displayed in the VRE and informed that the actual test will begin after that.
7. The instructor stays in close proximity of the participant during the duration of the test in case he or she has any additional questions.

The dataset consists of data from 166 unique participants and was collected during the course of six months [78]. The gathering of data has taken place, for the most part, at the two universities mentioned above, which resulted in a dataset with a majority of younger adults studying at higher level education. Eighteen additional participants, of similar background, have performed the test in addition to the ones from [78], using the same environment and instructions. This additional data collection took place in the office of the author of this thesis. The resulting dataset has 184 participants in total that answered 10 items each, which equals 1840 points in total. The age distribution for the dataset ranges from the youngest being 17 and the oldest 70 years old with an average age of 31. The gender division amongst the participants is 37.5% female, 62.5% male and 0% other.

7. Preprocessing and Feature Selection

The data from the tests was loaded into the computer memory from previous storage in files on the harddrive and the first item, Item 1, from each test participant was discarded as suggested by [78]. The reasoning behind this decision was, according to [78], that the participants spent a disproportionate amount of time on this item given its difficulty and

that this could be explained by the fact that most participants were unfamiliar with VR and ET prior to the test study.

It was determined, from visual inspection, that the dataset showed a similar structure as in **Paper A-C** and the same way of filtering outliers was therefore employed. The statistics of the dataset, before and after filtering, can be seen in Table 7.4. The threshold for the maximum length was set to 3100 based on that Mean + $3*\sigma = 806+3*763 = 3095 \approx 3100$, which is of reasonable magnitude since [78] shows that the average duration for the most difficult item (10) is 20s, i.e. roughly 2000 samples. The consequence of the filtering is that the dataset is slightly reduced from 1656 data points (N) to 1595 data points (N) and that the maximum length (Max) of a data point drastically decreases, from 20687 to 3047, as shown in Table 7.4.

Table 7.4: Distribution information for unfiltered and filtered data.

Type	Mean	σ	Min	Max	N
Unfiltered	806	763	96	20687	1656
Filtered	667	519	96	3047	1595

The features, Table 7.5, that were used in this work are the average gaze direction vector, obtained as the average vector from the left and the right eye, together with a variable, *Has6Alternatives, [-1 or 1]*, that tells the network if there are 6 [1] or 8 [-1] alternatives available for the participant to answer. This was added as a way to help the network identifying regions of interest after observing in the collected data that the alternatives are placed slightly differently depending on whether there are 6 or 8 alternatives present in the VRE. The alternatives 4-6 for the items with only six options was relabeled as 5-7, due to this misalignment, in order to remove the conflict between these labels and their relative positioning, in the VRE, compared to the ones with 8 alternatives.

Table 7.5: Description of data used in classification.

Type	Feature
Input	AvgEyeDirection [x, y, z]
Input	Has6Alternatives (-1 or 1)
Label	SelectedAlternative (1-8)

The next step was to filter out each sample, within each data point, that contained NaN values and fill it with the previous valid sample. These occur when the ET fails to read the eye properly, the most common is that they are obtained as a result of the participant blinking.

The data points that were of shorter length than the decided threshold (3100) were padded with zeros (ZP) at the end to guarantee the structure of the data point that is fed to the network. After ZP, the data was normalized featurewise in the range from -1 to 1. The

normalization makes sure that different values of magnitude between features do not bias the network to emphasize the importance of one feature over another.

To make the classification possible each data point is coupled with a label that describes which alternative (1-8) that was selected by the test participant.

Once the data was filtered, ZP:ed and normalized it was randomly split into three categories, training/validation/test. The proportions of the splits are: 45% of the data for training, 5% for validation, and the remaining 50% was used for testing and evaluation of the network.

8. Neural Network Design

The CNN used in the extension to **Paper D**, Fig. 7.6, has been built as follows; the network takes the matrix X as the input and feeds it to a slightly modified version of the feature extraction that was used in [87]. The difference being that the first layer of the feature extraction has increased *size* and *stride*, 25 each, that is used to significantly reduce the dimensions of the input. The other adjustment is that the AvgPooling1D that was followed by a flattening has been replaced with a single GlobalAvgPooling1D layer that achieves a similar outcome, as it was done in **Paper C**, extracting the most important features and reduces the dimensions of its input. The number of filters that were used at the different stages are also adjusted to fit the current data and objective.

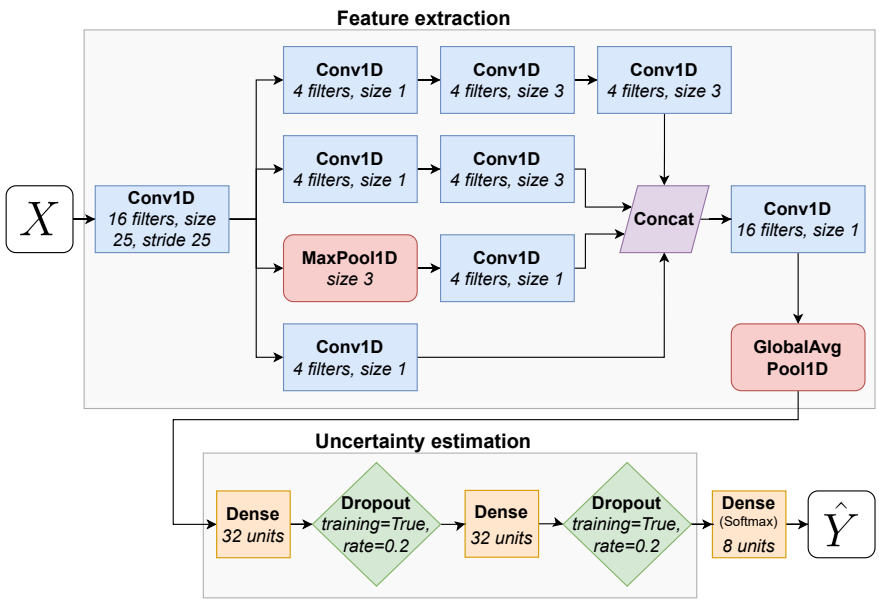


Figure 7.6: A flowchart that describes the network architecture used in the extension of **Paper D**.

9. Results

The classification results from this study, using the network described above, have been obtained using the same UE approach and filtering as described in **Study 1**, Section 7.1.

The classification results for different lower limits can be seen in Table 7.6. The results show that it possible to determine what alternative that a participant selected as the answer to a logical pattern from RPM. Depending on the desired outcome one may increase the classification accuracy by increasing the level of certainty (lower limit threshold) that is required to consider the prediction as valid. However, this comes at the cost of the network not being able to classify all samples.

The results, when compared to the ones obtained in **Study 1**, Section 7.1, shows that it is possible to use the same experimental procedure as described in Chapter 6 in two different contexts, namely for a movement based classification task and a decision based classification task from a psychological test.

Table 7.6: Comparison of classification results for different levels of filtering using UE and ZP.

Lower limit	% samples classified	Accuracy
0.00	100.00	80.33
0.10	94.36	82.74
0.20	90.98	84.30
0.30	86.34	85.63
0.40	82.71	86.97
0.50	79.20	87.82
0.60	74.56	89.41
0.70	69.92	90.32
0.80	61.40	91.84
0.90	46.37	94.05

CHAPTER 8

Summary of included papers

This chapter provides a summary of the included papers.

8.1 Paper A

Julius Pettersson and Petter Falkman
Human Movement Direction Classification using Virtual Reality and Eye Tracking
Published in Procedia Manufacturing, Volume 51, (pp. 95-102), 2020.

Combining the areas of virtual reality, eye-tracking and machine learning can be one way to increase the intelligence of collaborative robots. This could be broken down into the three stages, **Stage One:** *Movement Direction Classification*, **Stage Two:** *Movement Phase Classification*, and **Stage Three:** *Movement Intention Prediction*, described in the introduction. This paper gives a solution to the first stage and shows that it is possible to collect eye gaze data and use that to classify a person's movement direction. The results clearly shows that it is possible to combine virtual reality and eye tracking into a platform for testing and analysis of human behaviour, which can be beneficial in multiple areas of research. It is also shown that the implementation of uncertainty estimation improves the network and provides a way to improve the classification accuracy, at the cost of the percentage of samples classified, to obtain a more confident network.

8.2 Paper B

Julius Pettersson and Petter Falkman
Human Movement Direction Prediction using Virtual Reality and Eye Tracking
In 2021 22nd IEEE International Conference on Industrial Technology (ICIT), virtual, 2021.

One way of potentially improving the use of robots in a collaborative environment is through prediction of human intention that would give the robots insight into how the operators are about to behave. An important part of human behaviour is arm movement and this paper presents a method to predict the angular movement direction based on the human eye gaze. A test scenario has been designed in order to gather coordinate-based hand movement data in a virtual reality environment. The results shows that the eye gaze data can successfully be used to train an artificial neural network that is able to predict the direction of movement ~500ms ahead of time. It is also shown that a deeper and wider neural network does not necessarily always give better results.

8.3 Paper C

Julius Pettersson and Petter Falkman
Human Arm Movement Intention Prediction using Eye Tracking
Submitted to IEEE Transactions on Industrial Informatics.

Collaborative robots are becoming increasingly more popular in industries, providing flexibility and increased productivity for complex tasks. However, the robots are still not that interactive since they cannot yet interpret humans and adapt to their behaviour, mainly due to limited sensory input. Rapidly expanding research fields that could make collaborative robots smarter through an understanding of the operators intentions are; virtual reality, eye tracking, big data, and artificial intelligence. Prediction of human movement intentions could be one way to improve these robots. This paper shows that it is possible to collect gaze data and use that to continuously predict human movement intention on incoming data, utilizing the notion of uncertainty to determine whether a prediction should be trusted or not. The developed system reaches an accuracy of 70.7%, for predictions with high certainty, and an average of 37.2% ahead of the action being completed. Using the median length of a movement, this corresponds to 150 samples $* 0.372 \approx 550$ms.

8.4 Paper D

Julius Pettersson, Anton Albo, Johan Eriksson, Patrik Larsson, Kerstin W. Falkman, and Petter Falkman
Cognitive Ability Evaluation using Virtual Reality and Eye Tracking
In 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), (pp. 1-6), 2018.

The aim of this paper was to implement a simplified version of Raven's Progressive Matrices in a virtual reality environment where the eye gaze data is saved and compiled to

a document that can be used by psychologists during a diagnostic process or in research. The data has potential to indicate how the test persons reason while solving the different problems and could be implemented as an extension of the psychologists current toolbox. Furthermore, it has been found that test participants are less distracted by external disturbances due to the virtual environment implementation. The virtual environment could also be extended, as part of future work, to include external disturbances that could be manipulated by the test conductor in order to investigate how the different test participants respond to different types of disturbances.

CHAPTER 9

Discussion, Conclusions and Future Work

The ability to determine an upcoming action or what decision a human is about to take, can be useful in multiple areas, for example in manufacturing where humans working with collaborative robots, where knowing the intent of the operator could provide the robot with important information to help it navigate more safely. Another field that could benefit from a system that provides information regarding human intentions is the field of psychological testing where such a system could be used as a platform for new research or be one way to provide information in the diagnostic process. The work presented in this thesis investigates the potential use of virtual reality as a safe, customizable environment to collect gaze and movement data, eye tracking as the non-invasive system input that gives insight into the human mind, and deep machine learning as the tool that analyzes the data. The thesis defines an experimental procedure that can be used to construct a virtual reality based testing system that gathers gaze and movement data, carries out a test study to gather data from human participants, and implements an artificial neural network in order to analyze human behaviour. This is followed by four studies that gives evidence to the decisions that were made in the experimental procedure and shows the potential uses of such a system.

It is possible that the VREs that are used in the studies, Chapter 7, are contain biases. For example, the nature of the tasks that have been implemented guarantees that the participant has to direct the gaze towards interesting areas to solve most of the tasks. The randomness that is present in most of the tasks is also inhibiting a learning process that could potentially move participants from using the foveal vision to utilizing more of the periphery as, for example, the placement of products in a picking station is learnt. The goal of the simplicity of the tasks was, however, to enable a discussion regarding the performance of the different systems and the behaviour of the participants.

In Chapter 4, the concept of fixations and how these can be calculated was explained. From experimental results these showed little to no performance gain in the studies that have been performed. However, this should be investigated further since the lack of importance might either be due to that the network learns to calculate these on its own or that the tasks that was used are not complex enough to make the fixations provide any additional information to solve the objective. Finding, for example, a box that has been lit may mostly rely on peripheral vision, which excels at detecting changes in brightness whereas a more complex search task would probably rely more on using foveal vision to distinguish between objects or patterns. One other important aspect to consider is that a system based on eye gaze will always have limited abilities to analyze decisions made from peripheral vision since the ET hardware is only capable of measuring the direction of the foveal vision.

There are two types of arm movements that are gathered in **Paper B** and**Paper C** , namely one step movements, such as forward, backward, left, and right, along with sweeping movements to the left and right. The sweeps were, however, not considered in **Paper C** due to the fact that the gaze-hand connection was very different, the eye movements resemble smooth pursuits rather than fixation/saccade, to the other movements and did not fit the objective. These should be further analyzed in the future since there probably are other tasks and objectives where this behaviour comes in to play.

The test studies described in Chapter 7 are somewhat limited, in the way the data has been collected. In order to fully ensure that the methodology works on a more generalized scale one should redo these tests, or similar tests, with control groups that are larger and where the participants have been selected by people that are experts in creating diversified test groups.

There exists multiple other neural network types that potentially could be used to process the gaze data. One that has shown great promise in other fields that also deals with sequences of data is the Transformer architecture. It could be interesting to see if such a network perform differently compared to the approach that was used in **Paper C**.

RQ1: *Is is possible to predict human intention through the study of eye gaze?*

The work in **Paper B** showed that it is possible to predict the intended angular movement direction a fixed timestep ahead, however, with some difficulties to distinguish between smaller angles. These results were improved in **Paper C** where it was shown that the reaches an accuracy of 70.7%, for predictions with high certainty, and an average of 37.2% ahead of the action being completed, on a continuous stream of eye gaze data. Using the median length of a movement, this corresponds to 150 samples $* 0.372 \approx 550$ms.

RQ2: *Is DML a suitable tool to analyze the connection between eye gaze and intention in humans?*

In **Paper A-B**, it is shown that DML can be used to identify behavioural patterns in eye gaze data, both for classification of movements as well as prediction of angular movements. However, the complexity of the objective in **Paper C** is what really shows the capabilities of DML. The algorithm in **Paper C** continuously classifies the intended movement direction of the human, achieving 70.7% accuracy on its predictions through the use of the algorithm's notion of its own uncertainty.

RQ3: *How can a VRE-test be designed to gather the necessary eye gaze and movement data to be used in an application based on DML?*

Paper D showed that it is possible to use VR to gather eye gaze and movement data from humans performing the task of solving a logical pattern. This formed the basis for **Paper A-C**, where this concept is transferred from logical patterns to tasks involving reaching for specific objects in the VRE and the data is successfully used to train a DML algorithm. **Paper D** did not investigate the use of DML, however, the extension to this paper, described in Section 7.4, shows promising results.

9.1 Future Work

The VREs used to achieve the results in **Paper A-C** contains simplified tasks that are similar to the ones present in, for example, a pick-and-place station in a manufacturing environment. In order to further evaluate the described procedure and the results from **Paper C**, the implementation of a VRE with more complex tasks would be of interest. This could include an assembly station were the operator collaborates with a virtual robot with an external control system that receives the predicted intentions and makes the robot adapt accordingly.

A natural extension to the suggestion above, if the results are determined successful, would be to implement the system in a real world application, preferably similar to the industrial application that is used for the step above. This would include using safety glasses with built-in ET instead of a VR-headset and the evaluation of the validity of using such a system in a realtime environment.

Another topic to explore would be to implement a full version of RPM, gather data from a control group selected by researchers in the field of psychology, and together with them determine a few interesting objectives that may be solved using the procedure described in this thesis. This could potentially be to extend the neural network from **Study 4**, Section 7.4, to determine what alternative was the second most considered or to implement the functionality from **Paper C** to perform analysis on continuous data.

References

[1] I. El Makrini, K. Merckaert, D. Lefeber, and B. Vanderborght, "Design of a collaborative architecture for human-robot assembly tasks", in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 1624–1629.

[2] J. Krüger, T. K. Lien, and A. Verl, "Cooperation of human and machines in assembly lines", *CIRP annals*, vol. 58, no. 2, pp. 628–646, 2009.

[3] M. Awais and D. Henrich, "Human-robot collaboration by intention recognition using probabilistic state machines", in *19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010)*, IEEE, 2010, pp. 75–80.

[4] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration", in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, IEEE, 2016, pp. 83–90.

[5] L. Bi, C. Guan, *et al.*, "A review on emg-based motor intention prediction of continuous human upper limb motion for human-robot collaboration", *Biomedical Signal Processing and Control*, vol. 51, pp. 113–127, 2019.

[6] H. chaandar Ravichandar, A. Kumar, and A. Dani, "Bayesian human intention inference through multiple model filtering with gaze-based priors", in *2016 19th International Conference on Information Fusion (FUSION)*, IEEE, 2016, pp. 2296–2302.

[7] R. M. Bakwin, A. Weider, and H. Bakwin, "Mental testing in children", *The Journal of pediatrics*, vol. 33, no. 3, pp. 384–394, 1948.

[8] A.-C. Smedler and E. Tideman, *Att testa barn och ungdomar : om testmetoder i psykologiska utredningar*, 1. utg. Stockholm: Natur & kultur, 2009, ISBN: 978-91-27-11692-4 (inb.)

[9] C. Karatekin, "Eye tracking studies of normative and atypical development", *Developmental review*, vol. 27, no. 3, pp. 283–348, 2007.

[10] A. Poole and L. J. Ball, "Eye tracking in hci and usability research", *Encyclopedia of human computer interaction*, vol. 1, pp. 211–219, 2006.

[11] F. Jungwirth, M. Murauer, M. Haslgrübler, and A. Ferscha, "Eyes are different than hands: An analysis of gaze as input modality for industrial man-machine interactions", in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, ACM, 2018, pp. 303–310.

[12] L. Wu, L. Guo, H. Fang, and L. Mou, "Bullet graph versus gauges graph: Evaluation human information processing of industrial visualization based on eye-tracking methods", in *International Conference on Applied Human Factors and Ergonomics*, Springer, 2018, pp. 752–762.

[13] G. Tang, P. Webb, and J. Thrower, "The development and evaluation of robot light skin: A novel robot signalling system to improve communication in industrial human–robot collaboration", *Robotics and Computer-Integrated Manufacturing*, vol. 56, pp. 85–94, 2019.

[14] N. I. Vargas-Cuentas, D. Hidalgo, A. Roman-Gonzalez, M. Power, R. H. Gilman, and M. Zimic, "Diagnosis of autism using an eye tracking system", in *Global Humanitarian Technology Conference (GHTC), 2016*, IEEE, 2016, pp. 624–627.

[15] M. Dahl, A. Albo, J. Eriksson, J. Pettersson, and P. Falkman, "Virtual reality commissioning in production systems preparation", in *22nd IEEE International Conference on Emerging Technologies And Factory Automation, September 12-15, 2017, Limassol, Cyprus*, IEEE, 2017, pp. 1–7.

[16] A. A. Rizzo, M. Schultheis, K. A. Kerns, and C. Mateer, "Analysis of assets for virtual reality applications in neuropsychology", *Neuropsychological Rehabilitation*, vol. 14, no. 1-2, pp. 207–239, 2004.

[17] A. A. Rizzo, T. Bowerly, J. G. Buckwalter, D. Klimchuk, R. Mitura, and T. D. Parsons, "A virtual reality scenario for all seasons: The virtual classroom", *Cns Spectrums*, vol. 11, no. 1, pp. 35–44, 2009.

[18] M. Abidi, A. Al-Ahmari, A. El-Tamimi, S. Darwish, and A. Ahmad, "Development and evaluation of the virtual prototype of the first saudi arabian-designed car", *Computers*, vol. 5, no. 4, p. 26, 2016.

[19] A. M. Al-Ahmari, M. H. Abidi, A. Ahmad, and S. Darmoul, "Development of a virtual manufacturing assembly simulation system", *Advances in Mechanical Engineering*, vol. 8, no. 3, p. 1 687 814 016 639 824, 2016.

[20] D. Aschenbrenner, N. Maltry, J. Kimmel, M. Albert, J. Scharnagl, and K. Schilling, "Artab-using virtual and augmented reality methods for an improved situation awareness for telemaintenance", *IFAC-PapersOnLine*, vol. 49, no. 30, pp. 204–209, 2016.

[21] J. Pettersson, A. Albo, J. Eriksson, P. Larsson, K. Falkman, and P. Falkman, "Cognitive ability evaluation using virtual reality and eye tracking", in *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, IEEE, 2018, pp. 1–6.

[22] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data: The management revolution", *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

[23] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", *arXiv preprint arXiv:1708.08296*, 2017.

[24] K. Nagorny, P. Lima-Monteiro, J. Barata, and A. W. Colombo, "Big data analysis in smart manufacturing: A review", *International Journal of Communications, Network and System Sciences*, vol. 10, no. 3, pp. 31–58, 2017.

[25] O. Morariu, C. Morariu, T. Borangiu, and S. Răileanu, "Manufacturing systems at scale with big data streaming and online machine learning", in *Service Orientation in Holonic and Multi-Agent Manufacturing*, Springer, 2018, pp. 253–264.

[26] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications", *Journal of Manufacturing Systems*, vol. 48, pp. 144–156, 2018.

[27] Y. Cui, M. Gierl, and Q. Guo, "Statistical classification for cognitive diagnostic assessment: An artificial neural network approach", *Educational Psychology*, vol. 36, no. 6, pp. 1065–1082, 2016.

[28] G. Deshpande, P. Wang, D. Rangaprakash, and B. Wilamowski, "Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data", *IEEE transactions on cybernetics*, vol. 45, no. 12, pp. 2668–2679, 2015.

[29] Psykologförbundet, *Hantering och förvaring av psykologiska test inom hälso- och sjukvården*, [Online], `https://www.psykologforbundet.se/globalassets/omforbundet/hantering-och-forvaring-av-psykologiska-test.pdf`, 2013.

[30] J. H. Elder, "Videotaped behavioral observations: Enhancing validity and reliability", *Applied Nursing Research*, vol. 12, no. 4, pp. 206–209, 1999.

[31] R. Adams, P. Finn, E. Moes, K. Flannery, and A. Rizzo, "Distractibility in attention/deficit/hyperactivity disorder (adhd): The virtual reality classroom", *Child Neuropsychology*, vol. 15, no. 2, pp. 120–135, 2009.

[32] Y. Pollak, P. L. Weiss, A. A. Rizzo, M. Weizer, L. Shriki, R. S. Shalev, and V. Gross-Tsur, "The utility of a continuous performance test embedded in virtual reality in measuring adhd-related deficits", *Journal of Developmental & Behavioral Pediatrics*, vol. 30, no. 1, pp. 2–6, 2009.

[33] M. Dyck, M. Winbeck, S. Leiberg, Y. Chen, R. C. Gur, and K. Mathiak, "Recognition profile of emotions in natural and virtual faces", *PLoS One*, vol. 3, no. 11, e3628, 2008.

[34] P. Lindner, A. Miloff, W. Hamilton, L. Reuterskiöld, G. Andersson, M. B. Powers, and P. Carlbring, "Creating state of the art, next-generation virtual reality exposure therapies for anxiety disorders using consumer hardware platforms: Design considerations and future directions", *Cognitive Behaviour Therapy*, pp. 1–17, 2017.

[35] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze", *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[36] A. I. Goldman *et al.*, *Theory of mind*, 2012.

[37] A. Armanini and N. Conci, "Eye tracking as an accessible assistive tool", in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, IEEE, 2010, pp. 1–4.

[38] A. Navab, K. Gillespie-Lynch, S. P. Johnson, M. Sigman, and T. Hutman, "Eye-tracking as a measure of responsiveness to joint attention in infants at risk for autism", *Infancy*, vol. 17, no. 4, pp. 416–431, 2012.

[39] D. Riby and P. J. Hancock, "Looking at movies and cartoons: Eye-tracking evidence from williams syndrome and autism", *Journal of Intellectual Disability Research*, vol. 53, no. 2, pp. 169–181, 2009.

[40] P. Deans, L. O'Laughlin, B. Brubaker, N. Gay, and D. Krug, "Use of eye movement tracking in the differential diagnosis of attention deficit hyperactivity disorder (adhd) and reading disability", *Psychology*, vol. 1, no. 04, p. 238, 2010.

[41] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning", *Perspectives on Psychological Science*, vol. 12, no. 6, pp. 1100–1122, 2017.

[42] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior", in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE, vol. 2, 2005, pp. 568–573.

[43] J. A. Russell and J. M. Fernández-Dols, *The psychology of facial expression*. Cambridge university press, 1997.

[44] C. Chandler, T. B. Holmlund, P. W. Foltz, A. S. Cohen, and B. Elvevåg, "Extending the usefulness of the verbal memory test: The promise of machine learning", *Psychiatry Research*, vol. 297, p. 113 743, 2021.

[45] C. Spearman, ""general intelligence," objectively determined and measured", *The American Journal of Psychology*, vol. 15, no. 2, pp. 201–292, 1904, ISSN: 00029556.

[46] J. C. Raven *et al.*, *Raven's progressive matrices*. Oxford Psychologists Press Oxford, 1998.

[47] J. W. Kalat, *Introduction to psychology. pacific grove, ca: Brooks*, 1996.

[48] J. Raven, "The raven's progressive matrices: Change and stability over culture and time", *Cognitive Psychology*, vol. 41, no. 1, pp. 1–48, 2000, ISSN: 0010-0285.

[49] ——, "The raven's progressive matrices: Change and stability over culture and time", *Cognitive psychology*, vol. 41, no. 1, pp. 1–48, 2000.

[50] J. C. Raven, J. Court, and J. Raven, *Manual for Raven's Progressive Matrices and Vocabulary Scales by JC Raven, JH Court and J. Raven; Section2; Coloured Progressive Matrices*. Oxford Psychologist Press, 1995.

[51] J. Raven *et al.*, "Raven progressive matrices", in *Handbook of nonverbal assessment*, Springer, 2003, pp. 223–237.

[52] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The cave: Audio visual experience automatic virtual environment", *Commun. ACM*, vol. 35, no. 6, pp. 64–72, 1992, ISSN: 0001-0782.

[53] S. Choi, K. Jung, and S. Do Noh, "Virtual reality applications in manufacturing industries: Past research, present findings, and future directions", *Concurrent Engineering*, vol. 23, no. 1, p. 56, 2015.

[54] T. D. Gould, T. M. Bastain, M. E. Israel, D. W. Hommer, and F. X. Castellanos, "Altered performance on an ocular fixation task in attention-deficit/hyperactivity disorder", *Biological psychiatry*, vol. 50, no. 8, pp. 633–635, 2001.

[55] A. T. Duchowski and A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2017.

[56] R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal, "Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction", *Robotics and Computer-Integrated Manufacturing*, vol. 61, p. 101 830, 2020.

[57] J. Hartwig, A. Kretschmer-Trendowicz, J. Helmert, M. Jung, and S. Pannasch, "Revealing the dynamics of prospective memory processes in children with eye movements", *International Journal of Psychophysiology*, vol. 160, pp. 38–55, 2021.

[58] M. Hochman, Y. Parmet, and T. Oron-Gilad, "Pedestrians' understanding of a fully autonomous vehicle's intent to stop: A learning effect over time", *Frontiers in psychology*, vol. 11, 2020.

[59] V. K. Sharma, L. Murthy, K. Singh Saluja, V. Mollyn, G. Sharma, and P. Biswas, "Webcam controlled robotic arm for persons with ssmi", *Technology and Disability*, no. Preprint, pp. 1–19, 2020.

[60] A. Keshava, A. Aumeistere, K. Izdebski, and P. Konig, "Decoding task from oculomotor behavior in virtual reality", in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[61] W. R. Hendee and P. Wells, *The perception of visual information*, 1993.

[62] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction", in *Advances in physiological computing*, Springer, 2014, pp. 39–65.

[63] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols", in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71–78.

[64] A. Olsen and R. Matos, "Identifying parameter values for an i-vt fixation filter suitable for handling data sampled with various sampling frequencies", in *proceedings of the symposium on Eye tracking research and applications*, 2012, pp. 317–320.

[65] A. Olsen, "The tobii i-vt fixation filter", *Tobii Technology*, vol. 21, 2012.

[66] P. Langley, "The changing science of machine learning", *Machine Learning*, vol. 82, no. 3, pp. 275–279, 2011.

[67] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[68] A. K. Jain, "Data clustering: 50 years beyond k-means", *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[69] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[71] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation", *arXiv preprint arXiv:1406.1078*, 2014.

[72] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012.

[73] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning", in *international conference on machine learning*, 2016, pp. 1050–1059.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", *arXiv preprint arXiv:1706.03762*, 2017.

[75] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks", in *International Conference on Machine Learning*, PMLR, 2016, pp. 1747–1756.

[76] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio", *arXiv preprint arXiv:1609.03499*, 2016.

[77] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[78] T. Bergström and J. Pettersson, "Autonomous gathering and clustering of behavioural data in virtual reality", Master's thesis, 2018.

[79] D. W. Hansen and P. Majaranta, "Basics of camera-based gaze tracking", in *Gaze interaction and applications of eye tracking: Advances in assistive technologies*, IGI Global, 2012, pp. 21–26.

[80] M. L. Mele and S. Federici, "Gaze and eye-tracking solutions for psychological research", *Cognitive processing*, vol. 13, no. 1, pp. 261–265, 2012.

[81] L. Zhang, J. Sturm, D. Cremers, and D. Lee, "Real-time human motion tracking using multiple depth cameras", in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 2389–2395.

[82] A. H. Moreira, S. Queirós, J. Fonseca, P. L. Rodrigues, N. F. Rodrigues, and J. L. Vilaça, "Real-time hand tracking for rehabilitation and character animation", in *2014 IEEE 3nd International Conference on Serious Games and Applications for Health (SeGAH)*, IEEE, 2014, pp. 1–8.

[83] Tobii AB, *Tobii pro vr integration – based on htc vive development kit description*, v.1.7 - en-US, Accessed on: Feb. 13, 2020. [Online]. Available: `https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-vr-integration-product-description.pdf/?v=1.7`, Tobii AB.

[84] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[86] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[87] J. Pettersson and P. Falkman, "Human movement direction classification using virtual reality and eye tracking", *Procedia Manufacturing*, 2020.