M.Sc.Eng. Thesis
Master of Science in Engineering

DTU Compute
Department of Applied Mathematics and Computer Science

Master Thesis

Artificial Intelligence for Estimating People's Perception of the World

Supervisor: Thomas Bolander

Author: Emilie Isabella Dahl (s153762)

Sunday 14th February, 2021
Kongens Lyngby

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Abstract

Abstract - English

One important human factor for society to accept human-robot social interactions is circumstance-appropriateness. This can be implemented through theory of mind, which is an assessment of human's degree of capacity for empathy and understanding of others. Implementing theory of mind on a robot will allow the robot to understand how people's beliefs are subjective. For this implementation to know when a false belief is created, an estimation of people's perception of the world is needed. This thesis explores how people create their perception of the world and conducts an experiment based on this.

Perception is based on different types of attentional factors. It was discovered that only some of these factors are observable thus implementing a completely accurate estimation is not feasible. These attentional factors are all of the type overt attention, and consists among others of eye movements and saliency. Issues such as inattentional blindness, which is when people do not notice a stimuli when looking directly at it, are not observable and can therefore not be estimated.

An experiment is conducted testing three methods for estimating individuals' capability to perceive a single target. The methods all rely on the deep neural network Gaze360, which estimates gaze directions of people within the images. Gaze360 falls under the category Scene Analysis with a third-person view, as both the people and the scene are visible within the images. The best method resulted in an accuracy of 82.1% by using the quantile offset from Gaze360 to create a Von Mises probability distribution of the participant's field of vision. The target's edges were extended by 30 degrees to reflect that humans are able to perceive movements within their peripheral field of vision.

A small dataset is created to test the methods. It consists of 20 videos of approximately 16 seconds each. In 6 of the videos the participant always look at the target, and in another 6 videos the participant never looks at the target. Lastly, 8 of the videos show two participants interacting with the target.

Abstract - Danish

Forståelse for hvilken opførsel der er passende, er en vigtig egenskab robotter mangler, for at samfundet accepterer sociale interaktioner mellem mennesker og robotter. Dette kan implementeres vha. theory of mind, som er en vurdering af menneskers evne til at føle empati og forstå andres synspunkter. For at implementere theory of mind, er det nødvendigt at forstå, hvordan folks opfattelse af verden ser ud. Denne afhandling udforsker, hvordan folk danner deres opfattelse af verden og udfører et eksperiment baseret på denne teori.

Menneskers opfattelse af verden er baseret på forskellige opmærksomhedsfaktorer. Kun få af disse faktorer er observerbare, så en nøjagtig vurdering af menneskers forståelse af verden er ikke mulig. De opmærksomhedsfaktorer, som er observerbare kaldes overt opmærksomhed, og består bl.a. af øjenbevægelser og saliency. Scenarier såsom uopmærksom blindhed kan ikke observeres og kan derfor heller ikke estimeres.

Der er udført 3 eksperimenter som estimerer individers evne til at opfatte et enkelt objekt. Disse metoder er alle afhængige af det dybe neurale netværk Gaze360. Gaze360 falder ind under kategorien Scene Analysis ud fra en tredjepersons synsvinkel. Det estimerer derved folks blikretninger i et billede eller en video. Den bedste metode resulterede i en præcision på $82,1\%$ ved hjælp af kvartiler fra Gaze360. Kvartilerne er brugt til at modellere en Von Mises sandsynlighedsfordeling af deltagerens synsfelt. Objektets kanter, som bruges til at måle hvor meget af objektet folk kan se, er blevet udvidet med 30 grader for at afspejle, at mennesker er i stand til at opfatte bevægelser inden for deres synsfelts ydregrænser.

Et lille datasæt er lavet for at teste de forskellige metoder. Datasættet består af 20 videoer på 16 sekunder hver. I 6 af videoerne vil deltageren altid kigge på et bestemt objekt, mens i 6 andre videoer vil deltageren aldrig se på objektet. Derudover er der 8 videoer, som viser to deltagere der interagerer med objektet.

Preface

This thesis is based on the development of an artificial intelligence algorithm for estimating individual people's perception of their environment. The project is created in cooperation with the department of Applied Mathematics and Computer Science at the Technical University of Denmark. The thesis is created in fulfillment of the requirements for acquiring a Masters degree in Human-oriented Artificial Intelligence. The workload is 30 ECTS points, and the project has run from the 24th August 2020 to the 14th February 2021. The supervisor of the project is Thomas Bolander of DTU Compute.

Kongens Lyngby, Sunday 14th February, 2021

Emilie Isabella Dahl (s153762)

Acknowledgements

I would like acknowledge the following people for their guidance, cooperation and support throughout the process of this project.

Thank you to Thomas Bolander for guiding me throughout the process through weekly meetings and advice. Thank you to Lasse Dissing who helped discuss geometrical issues and introduced me to the systems used within their work with theory of mind.

In addition, I would like to thank Per Bækgaard, Dan Witzner Hansen, Fiona Bríd Mulvey and Andrea Dittadi who helped me understand the topic and gave good advice when I needed it.

Contents

CHAPTER 1

Introduction

The mind and cognitive competence of humans is said to be our main advantages compared to animals and machines [1]. In recent years, research within Human-Robot Interactions (HRI) have had a focus on discovering how we teach a machine to communicate and integrate with people and society. One of the main abilities a robot would need in order to be socially accepted, is being able to understand the individuality of people's perception of the world.

This thesis explores how humans construct their individual perspective and how it can be estimated through observable cues. An experiment attempts to estimate humans' perceptions of the world based on vision. Vision is emphasised as gaze tracking technologies can be used to gain an insight into what people see, and thus a partial understanding of their perception. The goal is thereby to improve Human-Robot Interactions by enabling robots to have knowledge of people's different world perceptions through gaze tracking technology.

1.1 Report overview

Initially, the need for this solution is emphasized in Chapter 2, which introduces different types of human-robot interactions and how they can relate to an understanding of individuals' belief system through theory of mind. How humans gain a perception of the world is then explored in Chapter 3, which emphasises the importance of the human attention and what attention actually is. Chapter 4 describes the more specific anatomy of the eyes and how information is transported to the brain. In addition, Chapter 4 also describes the strengths and limitations of the human's field of vision.

Chapter 5 and Chapter 6 are closely related. Chapter 5 introduces basic concepts within computer vision, specifically deep learning, which are used in the gaze tracking technologies described in Chapter 6. Chapter 6 explains the terminology and taxonomy of existing gaze tracking methods including examples and explanations on how each type relates to this study.

Based on these technologies, multiple methods for estimating whether an individual is able to perceive the status of a single target is implemented and described in Chapter 7. Chapter 8 discusses the results from the experiment and its relation to the work from the preceding chapters. Potential further improvements to the methods are highlighted in Chapter 9, whereafter the evaluation of the development process is explained in Chapter 10. Lastly, the overall work is concluded in Chapter 11 followed by appendices.

A link for the GitHub repository containing the implementation and the test results is in Appendix D.

CHAPTER 2

Robot's Conception of Human Belief

A robot's lack of understanding of social context has created great frustration for humans with a non-technical background working with robots in their daily lives [2]. For a robot to be socially accepted, the main trait it lacks is *circumstance-appropriateness*. Circumstance-appropriateness can be implemented using techniques modelling *theory of mind* on a social robot.

The following sections will describe the basics of Human-Robot Interaction, Section 2.1, as well as developing techniques for implementing theory of mind, Section 2.2.

2.1 Human-Robot Interaction

When designing a successful robot, one of the main factors to take into account is the human factors. Human factors is designing a product, process, or system with the intention of reducing human error, increasing productivity and enhancing safety and comfort [3]. How humans interact with the device is the core of human factors. In 2016 Thomas B. Sheridan, who is well respected in the human factors community, wrote an overview of the status and challenges associated with Human-Robot Interactions [4]. Over the years robots have evolved from the human-controlled master-slave servomechanisms, initially created for handling nuclear waste, to a broad range of applications with humans as supervisory control. Four of these evolving applications are described by Sheridan, where all have different levels of automation and supervision from humans.

1. **Telerobots**: Humans have supervisory control of robots in performance of routine tasks. An example of telerobots is robots doing assembly line tasks. People's relationship to these robots is typically planning, teaching, monitoring of automatic control, making repairs etc.
2. **Teleoperators**: Humans having remote control of vehicles in various hazardous or inaccessible environments for non-routine tasks. E.g. a robot designed to handle nuclear waste is a teleoperator.
3. **Automated Vehicles**: Vehicles guided by a human passenger, but operating using artificial intelligence. These vehicles are seen in highway and railway vehicles as well as commercial aircraft.
4. **Human–robot social interaction**: Humans directly interacting with a robot that provide entertainment, teaching, comfort and assistance particularly for children, elderly, autistic and handicapped people.

Each type of application has some challenges, particularly in relation to human factors. For telerobots, the greatest challenge is the safety of the robots, often seen as collision avoidance [4]. Teleoperators' main challenges are within the control interface, in particular provision of 360 degrees observation at the remote site and compensation of delays and dropout in communication. When continuous teleoperation is impossible in certain conditions, the robots can be reprogrammed by a human supervisor to execute pieces of an overall task automatically, thereby making it operate as a telerobot.

Automated Vehicles are shown to be able to be autonomous when guided by artificial intelligence in e.g. a self-driving car by Google [4]. However, it is exceedingly difficult for computer vision and artificial intelligence to learn the everyday social interactions between drivers. According to Sheridan, there is a demand for further research in the social aspects of driving and the degree of automation the vehicle can achieve safely. Commercial aircraft are highly autonomous as the operating pilots are mostly exerting supervisory control in guidance and navigation. The main challenge of all automated vehicles is actually the acceptance by the passengers [4].

The last application type of robots is human-robot social interactions. These interactions are the main component this paper is focused on. The human-robot social interactions are closely related to the idea of general-purpose robots, which are flexible for every-day tasks with non-technical humans. The topic of social interactions are thereby described in further details in Section 2.1.1.

2.1.1 Human–robot Social Interaction

The human factors needed for a social robot to be accepted by the general public is a popular research topic. Particularly interesting is the research into what qualities a robot would need to interact with hospital patients or the elderly so that they trust it e.g. as an exercise coach or delivering food trays.

These desired qualities have been analysed for logistics robots in e.g. office spaces [2]. A study conducted using a communicating robot in an office space concluded that *familiarity*, *circumstance-appropriateness*, and *social-role* were the most important issues to gain social acceptance. The familiarity of a robot depends on how relatable it looks and acts. Circumstance-appropriateness is an understanding of the given circumstances, e.g. when to interact with human and when not to. And finally, the social-role is highlighting a clear portrayal of the robot's purpose in the given environment.

(a) Robovie IV interacting with people at the office.

(b) Robovie IV carrying a bag with documents.

Figure 2.1: These images are from a robot-human interaction study conducted at an office [2].

The office space experiment's robot is shown in Figure 2.1, where it communicated with an individual, or carried objects for the workers as a delivery task. Throughout the experiment the authors identified three levels of interactions with the robot: Direct interaction, observing other's interaction and rarely seeing the robot at all. The level of interaction with the robot affected which issues (familiarity, circumstance-appropriateness, social-role) were important for acceptance of it. For the people who had direct interaction with the robot, the most important features were familiarity and social role. In comparison, people observing the robot or who had little to no relationship with it, the acceptance became highly dependent on circumstance-appropriateness. As the largest group of people have little to no contact with robots, ensuring that robots understand circumstance-appropriateness is the most important factor for gaining social acceptance [2].

Today, some robots have already become socially accepted by the community. These robots either have strong social-roles, such as the cleaning robot Roomba, or an entertaining children's toy, such as the robotic toy-dog AIBO [2]. However, these examples are all related to personal robots, which are bought by people similar to the people who had direct contact with the robot from the office space.

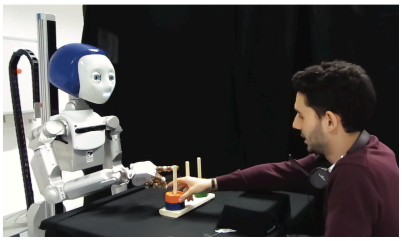
When designing a social robot for hospitals, offices, museums etc. they are no longer personal, and circumstance-appropriateness becomes increasingly important for everyone to accept it.

Additional work relating to how humans use social cues to interact with a social robot is described in Section 2.1.2

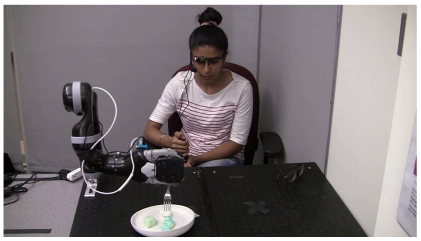
2.1.2 Teaching Robots Human Interaction

Various works describe how a robot's action influences humans ([5], [6]). One of the experiments show how designing a robot's gaze to imitate distinct personality types, e.g. introvert or extroverts, influences a user's motivation and rating of the robot [5]. Users were asked to solve a puzzle, which the robot had to teach and motivate them to keep solving. The test-setup is shown in Figure 2.2a. This work was mostly related to improving the familiarity of robots, as it would act based on individuals' personalities. The robot also has a strong social-role as the instructor, making its purpose clear to the user.

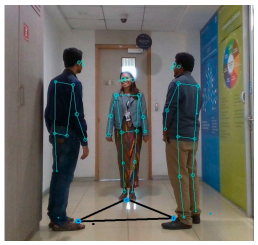
Another paper tested how human eyes and hand behaviors were influenced when sharing a manipulation task with a robot [6]. Unlike the previous experiment, this robot did not mimic the appearance of a human and was therefore less familiar. However, a strong sense of social-role allows the user to better understand its purpose and thus be more accepting of its presence. Figure 2.2b demonstrates how a participant used the robot-arm to pick up pieces of food from a plate. The participant is equipped with an eye tracker for monitoring eye movements in addition to the controls of the robotic arm. The work shows that the participants' pupils become larger when interacting with the robot. It seems that visible cues are noticeable in humans, to some extent, when interacting with a robot.



(a) A scene from [5], which analyses the motivation of people depending of the behavior of a robot.



(b) This image is from the work [6], which monitor eye and hand movements of people interacting with robots.



(c) A group in a triangular formation, where a robot is to calculate an appropriate angle to approach [7].

Figure 2.2: Examples of work related to human-robot social interactions. Each subfigure is from different papers within the field of HRI.

A final example of social robot experiments is teaching a robot to be socially aware of the conditions when approaching a group of people [7]. The intent for the robot is to be included in the conversation. Figure 2.2c shows an example of a standard group formation, triangular formation, where the robot needs to know how to join the group. This research allows the robot to become more aware of the circumstances, as it knows that it needs to join the group from the correct angle to be able to make the people aware and accept its presence.

2.2 Theory of Mind

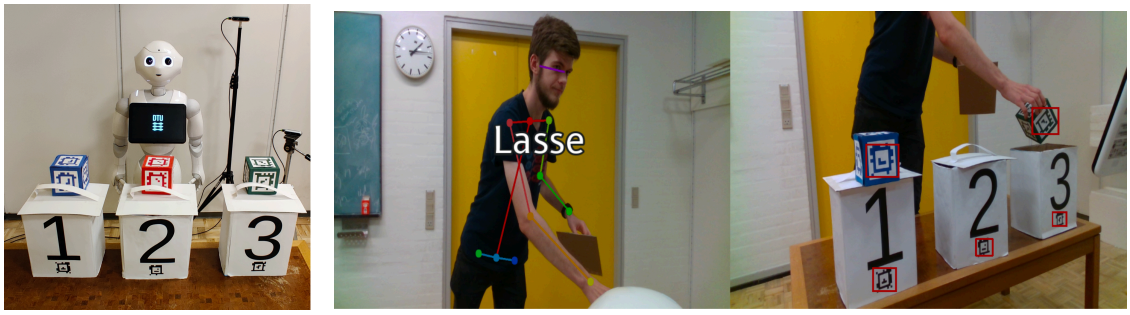
The work mentioned in Section 2.1.2 relates to specific tasks; motivating a person by mirroring their personality, reading the person's visible cues toward a robot, and understanding how to approach a social group. The work respectively focuses on familiarity, social-role and circumstance-appropriateness.

In order to design an interface with general social knowledge instead of task specific knowledge, it is essential to know the basic philosophy of how humans understand each other.

A study that explores solutions for improving human-robot interaction is a paper by Dissing and Bolander called *Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic* [8]. The focus of the paper is to teach robots theory of mind. Theory of mind is an individual's ability to understand and empathize with others through their desires, intentions and beliefs. Modelling the belief-system, in particular, can teach a robot to understand how individuals' knowledge differentiates, and thus be able to consider their circumstances. A method to test individuals' theory of mind is through *false-belief tasks* [9].

A general example of a false-belief task is the Sally-Anne test, which tells a story of two girls, Sally and Anne, who have a box, a basket and a marble [8]. Sally places the marble in the basket and leaves the room. Anne moves the marble to the box without Sally being present. The individual being tested will be asked *Where does Sally think the marble is?* If the subject correctly understands that Sally has a false belief of the marble still being in the basket, the subject has correctly understood a first-order false-belief task. In comparison, if subjects rely on real world truths and answer that Sally believes the marble to be in the box, they do not have theory of mind. Theory of mind is generally developed over time, and people learn it around the age of 6 [9]. Some people develop theory of mind more slowly or perhaps not at all. Children with autism have trouble understanding false-belief tasks, and might reach a mental age of about 10 years before they achieve an understanding [10]. Some complex problems similar to false-belief tasks, white lies or double bluff, can take them even longer to learn, and some may never fully grasp the concepts.

A second-order false-belief task is if Sally was actually spying on Anne, and thereby knows that the marble is moved. Now the subject will, instead, be asked *Where does Anne believe that Sally believes the marble is?*, as it is now Anne who has a false-belief of what Sally believes. These false-belief tasks can be continued in n-ordered false-belief tasks, e.g. if Anne secretly had a camera monitoring Sally, and thereby knew that Sally was spying on her. This would again result in Sally having a false-belief of Anne having a false-belief, etc.



(a) A SoftBank Robotics Pepper robot interpreting participant's beliefs.

(b) The data received from the two cameras recording the interactions.

Figure 2.3: The setup for the false-belief task created by Dissing et al. [8].

To test a robot's theory of mind, a task similar to the Sally-Anne test was created by Dissing et al.[8]. Figure 2.3a shows how a Pepper robot is stationed behind three numbered containers with small colored cubes on top. The robot keeps track of where people believe the cubes are located using cameras stationed around the robot, as seen in Figure 2.3b. The robot's task is similarl to an observer keeping track of Sally and Anne's beliefs in the Sally-Anne test. The false-belief task could in principle be conducted simply using a computer, a microphone and two cameras, but having the robot embodying an observing agent facilitates a better interaction for the participants [8].

In the given solution *Dynamic Epistemic Logic* (DEL) is used to teach the robot theory of mind.

DEL is a logical language that is used to model information changing events. This includes a model of the changes of the agents' believes and the actual real world changes [11]. In the proposed DEL solution the changing event are called using triggers. The solution therefore has multiple triggers e.g. when a lid of a container is lifted, a cube becomes visible, a cube is put into a container, a person leaves or enters the room, etc. The current algorithm assumes that when a person is visible and facing the camera, the person will be able to see the events occurring with the containers, thereby simplifying the estimation of what agents see in order to estimate what they believe. The algorithm that Dissing et al. created relies on the individuals attention always being on the containers and cubes, as well as not looking behind the others' back [8]. This is one of the shortcomings the work has in order to estimate people's perception of the world, and thus their beliefs. The importance of understanding human perception through gaze direction is highlighted in the work of Langton et al. [12] which is further described in Section 2.2.1.

2.2.1 Gaze Direction Relation to ToM

Various research within perception and attention is described in detail in Section 3, however, explicit theories are developed explaining the relation of human attention and theory of mind. Langton et al. described in February 2000 a number of different works involving social cues, theory of mind and gaze direction [12].

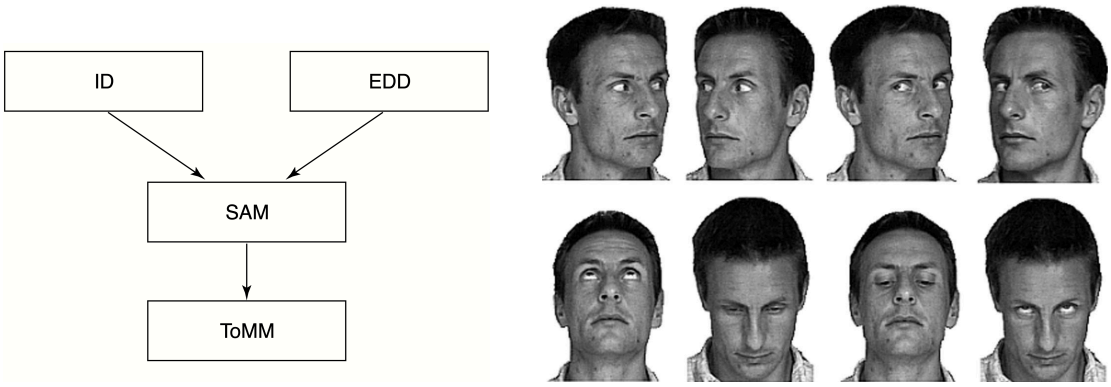
One of the theories presented by Langton et al. is by Perrett and his colleagues. Perrett et al. suggest that eye movements are used to regulate turn-taking in conversations, express intimacy and exercising social control [12]. These assumptions are also a basis for the research described in Section 2.1.2, where the authors try to understand whether human social cues are visible in human-robot interaction. Overall, the suggestion highlights the importance of human gaze for determining social cues and individual perceptions.

A concrete theory combining theory of mind with gaze direction is presented by Baron-Cohen as shown in Figure 2.4a [12]. Baron-Cohen proposes a *mindreading* model consisting of a collection of modules which allow humans to attribute the mental states of others. The model has 4 components:

1. **Intentionality Detector (ID)**: The ID is a primitive perceptual mechanism which interprets stimuli in terms of its desires and goals. So ID detector is a method to understand the goals and motivation of others.
2. **Eye-Direction Detector (EDD)**: The EDD simply detects eye-like stimuli and computes the gaze direction and point of regard.
3. **Shared-Attention Mechanism (SAM)**: SAM enables individuals to understand when a joint connection with another individual is created. Based on information from ID and EDD it understands the concept of having a joint connection when another agent's eyes are directed towards itself or the same location as itself.
4. **Theory-of-Mind Mechanism (ToMM)**: ToMM is the last mechnism in the model. It interprets SAM to create an understanding others intentions, believes and desires.

Therefore, according to Baron-Cohen, the use of gaze direction to establish joint attention underpins the development of a theory of mind. The model's specialized gaze module, EDD, is devoted to the task of detecting eyes and computing the point of regard in the environment. Such a devoted module would allow for rapid and obligatorily processing of the gaze of others.

Work mentioned by Langton et al. includes the search for a specialized neural circuitry tailored to perceiving others' gaze. An experiment by Perrett et al. identified certain cells in a macaque's brain, specifically the superior temporal sulcus (STS) of the temporal lobe, which responded maximally to a particular direction in which another monkey's eyes were looking. In addition, when the STS was missing, the monkeys were no longer able to make gaze-direction judgements. This indicates that



(a) A mindreading model suggested by Baron-Cohen [12].

(b) The test data for detecting whether eye direction or head orientation influence people's reaction time.

Figure 2.4: Figures from the research and work by Langton et al. [12].

a specialized cognitive module might exist for gaze tracking. The STS has later been claimed as a dedicated area in the brain within different cognitive functions as ToM, audiovisual integration, motion processing, speech processing, and face processing [13]. One paper proposes a multi-functionality of the STS region by having flexible network coactivations with other brain regions. So, the functionality of the STS become dependent on other regions of the brain [13].

Even if STS is not exclusively used for perceiving gaze direction, other studies have shown that humans react reflexively on gaze directions, which indicates a module similar to the EDD exist. These tests have shown that humans will react to gaze cues from other people reflexively even when told to ignore them [12]. These reactions occurred most often when the participants see the visual cues from their peripheral visual fields. In comparison, these reactions were not noticed at all when replacing the human gaze direction with e.g. an arrow.

Perrett and Emery suggested altering the model created by Baron-Cohen by replacing the EDD with a DAD, direction-of-attention detector. DAD was originally suggested as a hierarchy of attention indicators. The eyes were the most influential, but if the eyes were not visible the orientation of the head would be the next source of the individual's attention direction, and lastly the body posture.

Like Perrett and Emery, Langton et al. questioned whether gaze direction was the only indicator of a human's attention and source of processing [12]. They created an experiment where people had to judge where a person was looking based on either the head orientation or gaze direction. Examples of the images the subjects judged are shown in Figure 2.4b. The person in the images would be looking to the left, while the head was oriented to the right to confuse the participants. The reaction times were measured when determining the direction of attention based on eyes or head separately. If the main source of determined attention direction is the gaze, then the participants would be faster in determining the direction based on the eyes and not be influences by the head's orientation. The results showed that people were significantly faster when the eyes and head had the same direction, but no evidence suggested that the eyes were more important than the orientation of the head. Thereby, the orientation of the head has a parallel effect on reflectively understanding another human's attention [12]. Additional experiments found that the posture of the body and even gestures such as pointing also had similar effect. In addition, the same brain cells responding to gaze direction were also found to be sensitive to conjunctions of eye, head and body positions [12].

Thus, the DAD module suggested by Perrett and Emery might not fully grasp how attention shifts. However, it does include the influence of multiple features helping people communicate. Baron-Cohen's model seems significant as gaze has been proven to be able to indicate where other people's attention is directed, but the model is simplified. The model focus only on gaze direction and do thereby not

include additional features which also direct people's attention [12].

2.3 Summary

A potential solution to the problem of a robot's lack of circumstance-appropriateness is implementing the ability to understand a human's different beliefs. People who can interpret other's beliefs and understand that they might be different from reality has a theory of mind. To allow robots the same attributes, implementing a theory of mind on a robot seems as the first step. A suggestion as to how theory of mind is coupled to people's basic attention flow is shown in a model by Baron-Cohen, where one of the first sources of information is the human gaze. Langton et al. showed that head orientation and body posture are as important as the human gaze when determining the direction of a person's attention. So, to accurately implement theory of mind on a robot, it is essential that the robot is able to interpret people's perception of the world, and thus their head orientation and gaze direction. Understanding how people perceive the world is discussed in Chapter 3, while the technology within determining gaze direction is explained in Chapter 6.

CHAPTER 3

Perspective of the World

How do humans gain their perception of the world? This question is essential when designing a system which is able to interpret people's understanding of the state of the world. Studies show ([14]) that we do not always perceive what is right in front of us, as we might be focused on other tasks. This section describes how studies evaluate people's perception and attention, as well as highlight the potential shortcomings of a computational system based on visible cues from the human body. The overall question is whether it is even possible to create an accurate model given the current knowledge of how human's perception of the world works.

3.1 Cognitive Attention

According to research noted in *From Human Attention to Computational Attention: A Multidisciplinary Approach* by Mancas et al. *"Attention is the first step of perception"* [15]. The initial gateway of understanding how humans perceive the outside world is through the human senses: vision, audition, touch, smell or taste. Attention is the method which humans use to filter and analyse the outer real world and turn it into an inner conscious representation. However, the study of attention proves to be more complex than a simple *information reduction* system.

Attention was first mentioned in philosophy in 1890 by William James, known as the father of psychology. James wrote that *Everyone knows what attention is*, as the explanation for why the topic had not been researched before, it seems so fundamental. However, after almost 100 years, this statement was challenged by H. Pashler. Pashler spent several decades of research in cognitive psychology and claims that *No one knows what attention is* [15]. The study of attention had been divided and analysed in different fields including Philosophy, Experimental Psychology, Cognitive Psychology, Cognitive Neuroscience, and Computer Science. Mancas et al. describes the fields as the layers of an "attention onion" [15]. Philosophy was the first field of study to gain an interest, which sparked the interest respectively in the other fields. Each field of study has built different sets of knowledge on top of each other and divided the attentional concept into multiple layers.

The overarching question of "What is attention" is thereby quite complex and Mancas et al. writes that currently this question has no final answer. However, multiple kinds of attention have been established, which describe different methods of how the brain is provided with the capacity of selecting relevant information and prioritizing tasks. In addition, attention not only serves to select a location of interest, but also enhances the cortical representation of objects at that location, thus casting the location in a form of spotlight [16]. Attention can be seen as three concepts: eye movement, memory and action, and cognitive load. These concepts will be described in the following sections respectively.

3.2 Eye movements and Vision

The primary question to be answered by eye movements and vision within attention is *"where" the objects of interest are located.* This section will first introduce how the location of attention is not always observable. This is followed by an introduction to research within eye movement and how attention influences it.

3.2.1 Observable Attention

The methods analysed focuses on changes in people's posture, which can indicate that a person is preparing sensory receptors for an expected input. These changes could include eye movements, head movements, external ear movements, changes in pupil size etc.[15]. The observable changes are all examples of a type of attention called *overt* attention [15]. The contrast to overt attention is *covert* attention, where no posture changes are visible. The covert attention focuses on bringing regions of a scene to consciousness that is not fixated on by the eyes.

To estimate people's perception of the world, this study relies on physical movements which are visible in images, as this is the main indication of what they experience and thus what they see at the given time. Similarly, the study of human eye movement and vision is built on the visible indication of where humans have their attention. Eye movement alone is not the full and accurate depiction of where the person's attention is located as the other senses also contribute. However, eye movement is often depicted as one of the most important senses, and can give an insight into one of the attention's input channels. Defining the human senses in a hierarchy is challenged by Fabian Hutmacher, as he explains that the elaborate research within vision is mainly a cause of culture and current technology [17]. Nevertheless, the visual sense is visible from an image and thus available to analyse in this study, which is the main reason for the emphasis.

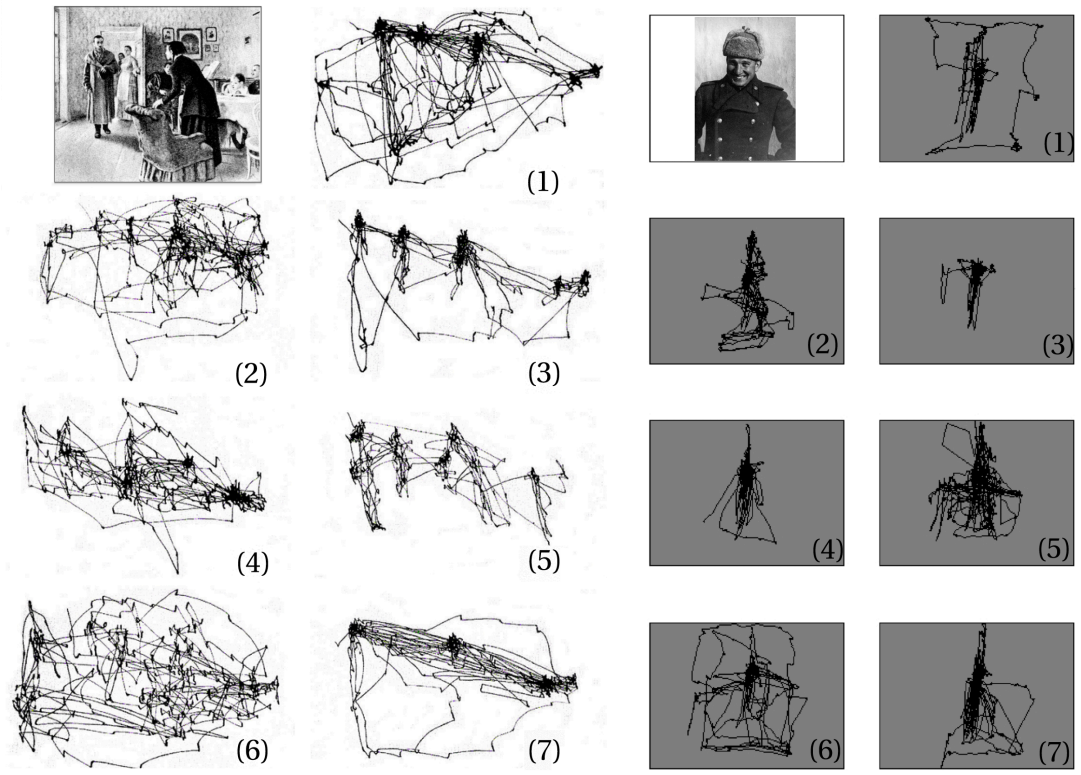
3.2.2 Founders of Modern Eye Movement Research

Al'fred Luk'yanovich Yarbus is known as one of the founders of modern eye movement research for his work in 1965 [18]. Yarbus wrote about stabilised retinal images, as well as cognitive influences on scanning patterns. Scanning patterns are shown to be a systematic method of how perspective of a scene effects eye movements, and therefore the areas attended to. Tatler et al. extened Yarbus' original work on cognitive scanning patterns and how prior instructions influence the way people look at the world [18].

Figure 3.1 shows the result of two experiments on human scanning patterns. The figure shows the original work of Yarbus, Figure 3.1a, as well as the extended work by Tatler et al., Figure 3.1b. The tests were conducted by having a number of subjects view the same picture multiple times with different instructions. The picture in Yarbus' original experience was a painting of a social scene called *The Unexpected Visitor* from 1888 by Ilya Repin. The image chosen by Tatler et al. was a portrait of Yarbus himself, showing a single man instead of a complex social scene. The intention of the second experiment was to challenge the concept of scanning patterns when looking at a close-up image of the face. A systematic cyclic pattern of looking between the eyes and the mouth has been established by Yarbus in a different chapter. The second experiment wanted to see if this pattern was noticeable when looking at a portrait of a person with his upper body included [18].

The location of the individual's eye movements were recorded over time illustrating the individual's scanning patterns. The instructions given to the subjects are described in Table 3.1. Both of the scanning patterns shown in Figure 3.1 are from a random person within the group of test subjects. The individual subjects did not have the same scanning patterns given the same instructions, but a significant difference was noticeable when comparing the individuals' own scanning patterns. The conclusion of Yarbus' paper was that people have different eye movements depending on the condition and task they were given. Thus, scanning patterns are influenced by the perspective of the scene.

The second experiment concluded that the facial features, especially the eyes and mouth, did receive the greatest fraction of the gaze allocation. However, the cyclic scanning between eyes and mouth was less evident, and not shown for some of the subjects. Thus, when an image contains both a face and part of the body the cyclic scanning pattern is not always visible. This could indicate that we view images and the real world differently, as the real world view will seldom only have the face visible [18]. The way we perceive the world is **task-dependent** and thus influence our eye movements as our attention shifts.



(a) The results from Yarbus' research (b) The results from Tatler et al.'s research

Figure 3.1: The divers scanning patterns of different images given instructions seen in Table 3.1 from [18]

Yarbus' Instructions	**Tatler et al.'s Instructions**
1. Free examination	1. Free examination
2. Estimate the material circumstances of the family in the picture	2. Estimate how wealthy this person was and what his social position was
3. Give the ages of the people in the picture	3. Estimate how old the person was when the picture was taken
4. Surmise what the family had been doing before the arrival of the 'unexpected visitor'	4. Estimate what the person had been doing just before this picture was taken
5. Remember the clothes worn by the people	5. Remember as much as you can about the clothes
6. Remember the position of the people and objects in the room	6. Try to remember the positions and details of everything
7. Estimate how long the unexpected visitor had been away from the family	7. Try to estimate how long this person had been away from home and why he had been away

Table 3.1: The instructions by Yarbus and Tatler et al. [18]

The original work by Yarbus' main focus was actually a method of creating stabilised retinal images, but the cognitive factors of eye movement described in the same paper have become increasingly popular. The cognitive factors were especially highlighted in 1990 through additional publications that related to the cognitive work of Yarbus [18].

3.2.3 Types of Eye Movements

To understand how the perspective of a scene influences eye movements, it is also important to understand which types of eye movements are actually possible. Today, three types of eye movements have been established: fixations, saccades and smooth pursuit. Fixation is when the eyes rest on a seemingly single location for a minimal period of time. Even though the eye might seem to look still, it has small micro-saccades, which simply means that the eye is never truly still. The second type is saccades, which are rapid eye movements between two locations (fixations). Often no visual data is acquired at the time of saccades, as the focus is to create a new fixation at another location. The final type of movement is a smooth pursuit. This is almost like a fixation, but when the fixation is located on a moving target. This will make the eyes follow the moving target in a slow pace.

The scanning patterns from Section 3.2.2 are based on where the different fixational point are located when viewing an image. The saccades do not jump too much, as the focus is mostly within the picture's frame. These days, eye technology often record fixation durations, saccade amplitudes and pupil size using an eye tracker.

3.3 Emotion and Action

The second concept within attention is emotion and action, which asks the question *"what" the object of interest is* rather than "where" as seen in the previous section. Research shows that people's attention is guided by bottom-up and top-down saliency [15]. These concepts highlight which objects might draw the attention of people. The concept of saliency is described in Section 3.3.1, which is followed by an example of visual search which also use saliency in Section 3.3.2.

3.3.1 Saliency

The two types of saliency are bottom-up and top-down. First, bottom-up saliency refers to exogenous, outside-the-person control of attention [15]. In images bottom-up saliency is based on image-based cues, which might be a bright, dense, weirdly shaped or very colorful area within the scene [16]. In other word, an area that stands out from the main format of the image. An example could be a white image with a single circle drawn on it. As the circle stands out within the white background, it will have a very high bottom-up saliency rate in contrast to the rest of the white background. For a bottom-up saliency example within sound, this might be a simple knock on the door in a silent house. You cannot help notice the knocking, as it stands out from quiet. Bottom-up saliency is not always a voluntary action, as the attention is sometimes caused by reflex. This reflex was necessary in the past, when an abnormality might be caused by potential danger.

In contrast to bottom-up saliency is top-down saliency, which refers to the endogenous or internal control of attention [15]. Unlike bottom-up, which is approximately the same for every living being (generalizable), top-down is different for every individual as it depends on subjective emotion and experiences (ungeneralizable). Top-down is thereby based on task-dependent cues related to memory, emotions and individual goals [16]. This makes it more complex for computers to model, as it requires information about the individual's internal states, goals, prior knowledge and/or emotions. Examples of such implementations of saliency, both bottom-up and top-down, are described in Section 6.2.2. The implementation of top-down saliency is often simplified as high-level features such as recognizing people, faces, hands, animals, object, etc. Emotions are often tied to meaningful targets, so the simplification

of subjective emotions tend to show good results. The first conceptual model of saliency was introduced by Koch and Ullman in 1985, and later implemented and refined by Itti and Koch [16].

In addition, top-down saliency can actually be divided into two sub-components: Goal/action related and Memory/emotion related [15]. The goal/action related top-down attention, also known as volitional top-down attention, is when the attention is focused on objects related to the current goals of the individual. When the attention is related to a goal, the user will not only react to novel situations, but also manage the goals and actions accordingly. So, volitional attention can be seen as a behavioral update when focusing on current actions to reach a desired goal.

In comparison, Memory/emotion related top-down attention is a process which is influenced by experience and prior knowledge. If you recognize a friend or see a spider, which you are terrified of, it will draw your attention. It does not have to be image related, you might simply hear a song you have fond memories of, which will thereby also catch your attention.

Finally, the different types of attentional concepts can be independent, but they often relate to each other. If you compare the three attentional concepts: bottom-up, volitional top-down, and memory/e-motion top-down, they can be seen in a sort of hierarchy. The bottom-up attention is the strongest and can pull your attention from either types of top-down attentions. Then volitional top-down attention can overtake attention from a memory-related top-down attention, which in turn is at the bottom. However, even though some types of attention are more likely to inhibit the other components of attention it is a balance that depends on the level of intensity. A good example is described by Mancas et al.:

> "If someone searches for his keys (volitional top-down), he will not take care about a car passing by. But if he hears a strange sound (bottom-up) and then recognizes a lion (memory-related top-down attention), he will stop searching for the keys and run away." [15]

3.3.2 Visual search

Visual search has become quite popular when analysing perspectives of a scene. Visual search is research in how human visually search scenes for targets. As described in Section 3.3.1, a person's attention is mostly influenced by bottom-up or top-down attention. A paper by Jeremy M. Wolfe describes how it is simply not possible to fully process all stimuli in a visual field at once, which is the reason for an attentional filter [19].

Wolfe describes how the first step of a search is the *pre-attentive attributes* within a scene, which might indicate where the target object is. The pre-attentive attributes are a limited set of attributes such as color, motion, size etc. that can be processed across the visual field in parallel [19]. An example is if you are given the task to find a blueberry. The pre-attentive attribute would be the color blue and have a small rounded shape. This would allow you to focus on the limited set of attributes within the scene and attain an overview across the whole field in a single step. The bottom-up attention is thus used to recieve an initial guidance through the image using the pre-attentive attributes in a stimulus-driven manner. The next step is top-down attention in a user-driven manner. The top-down attention allows the individual to recognize objects and thus recognize whether the desired target is at a given location.

As the individuals looks at a scene it is known that the target is not at places already searched. We therefore tend to ignore all areas already explored, which is also known as the *inhibition of return*. This allows for a faster search, as the highly salient areas are not explored twice. In addition, an oversight seen frequently when searching is the *satisfaction of search*. If we find the desired target, we are satisfied and move on. We ignore the possibility that there might be multiple identical targets as we are satisfied when simply finding one.

3.4 Cognitive Load

The final concept widely researched is the cognitive load. The cognitive load does not focus on answering a specific question, but evaluates how the capacity of attention is divided. The brain has a limited capacity, where five types of division of attention have been established. These five types are also known as the *clinical model of attention* [15]. The five variations are listed below.

1. **Focused attention**: When all attention is on a specific stimuli, thereby being focused on a precise task.
2. **Sustained attention**: When maintaining a consistent response during a longer continuous activity. So one might attain being attentive for a long period of time on the same topic.
3. **Selective attention**: When selectively maintaining cognitive resource on a specific stimuli and thereby ignoring potential distractors.
4. **Alternating attention**: When switching between multiple tasks, allowing the attention to focus on different tasks in a sequential process.
5. **Divided attention**: When dealing simultaneously with multiple tasks. This might also be known as multitasking, where the attention divided in a parallel process.

The five kinds of attention are all relevant, as the degree of focus influences the perception the individual will create. An interesting example is the study of selective attention, where one can become blind to otherwise obvious situations. This example is described in Section 3.4.1. The clinical model of attention is relevant as it distinguishes how clear a perspective the individual has of the real outer world. If the degree of focus is small, i.e. when one has divided attention, the perception of the world might not be as accurate, as if the attention is focused.

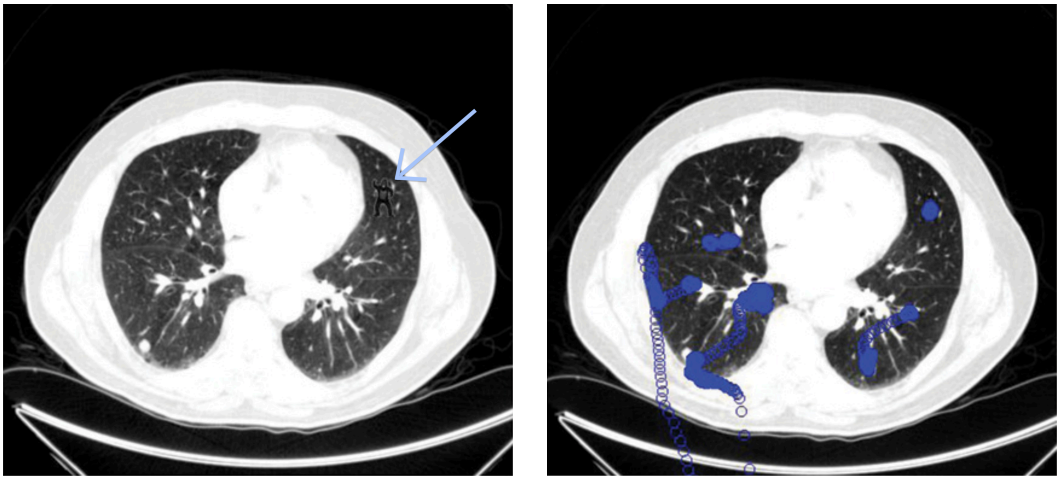
3.4.1 Selective Attention

Drew et al. conducted a study on inattentional blindness (IB), which is when people can look directly at a salient stimuli without noticing it [14]. Inattentional blindness is a concequence of selective attention, as a potential important stimuli is viewed as a distractor and thereby ignored.

This proves that even when implementing the best gaze tracking technology, there is a chance that the person does not actually perceive the stimuli right in front of them. The study of inattentional blindness is based of the well known study *the invisible gorilla* from 1999 by Simons and Chabris [14]. The observers are told to count a ball-passing game while a human in a gorilla suit wanders through the field of players. Even though the gorilla walked through the center of the scene, a substantial portion of the observers did not report seeing it.

Drew et al. extended this study to understand whether experts would notice a similar abnormality within their field of expertise [14]. The task was given to expert radiologists as well as novices, to spot nodules within five CT scans of the lungs. The last CT scan contains a hidden gorilla, which can be seen in Figure 3.2a. Only 16.7% of the radiologists located the gorilla, and none of the novices reported seeing it. Figure 3.2b shows the location of a radiologists who did not notice the gorilla, but evidently looked directly at it.

Given the high score for inattentional blindness Drew et al. did an additional test to ensure that the gorilla was actual visible [14]. So they showed a movie with accurate CT frames and ones containing the gorilla to a number of subjects. The result showed that the gorilla was correctly located 88% of the time, even with a high frame rate. In conclusion, it shows that expertise does have an effect on inattentional blindness, as a small percentage of the radiologists did see the gorilla. It is believed that the more difficult the task, the higher the inattentional blindness rate, which also explains why people who are used to looking at CT scans might find the task easier and thus become less blinded. However,



(a) Gorilla visible in CT scan.

(b) Eye tracking of a radiologist looking at the image.

Figure 3.2: Two identical CT scans highlighting a Gorilla in the left image, and the eye fixations in the right image. The images are from [14].

more important is the fact that the radiologist were able to detect lung nodules with a mean detection rate of 55% in comparison to the naive observers with a mean detection rate of 12%. In conclusion, even this high level of expertise does not immunize individuals against inherent limitations of human attention and perception. A better understanding of these limits could reduce the consequences of them so that medical and other man-made search tasks become more accurate [14].

3.5 Summary

In conclusion, there are three concepts relevant to gaining a perception of the world showing how attention is an essential part of estimating a person's consciousness. These concepts try to answer three different questions: "where" the objects of interest are located (Eye movement), "what" the objects of interest are (Emotin and Action), and how attention can be divided given a degree of focus (cognitive load). These concepts show that following visual cues from a human body will only allow an interpretation of the overt attention, thus ignoring convert attention. In addition, the limited ability for generalisation of top-down saliency creates a complex problem, as it seems unattainable to configure individuals' prior knowledge and emotions. Given that the scenario of selective attention creates the issue that a subject might be focusing on a different task and thereby experiencing inattentional blindness. Thus, given the current knowledge and technology available, it is not possible to create a completely accurate model of a person's perception of a scene. However, it is possible to estimate a perspective of the scene by relying on overt attention signals, scene estimations of bottom-up and top-down saliency, and scoping to focused, sustainable or alternating degrees of focus. The limitations are a shortcoming of an implementation, but an estimation of humans' perception of the world is an advantage to having none.

CHAPTER 4

Field of Vision

It is evident that attention, and thus our perception, cannot be be estimated accurately solely based on eye movements visible in an image (described in Chapter 3). Attention is mainly seen as a filtration of our senses, which allows us to process and create our perception of the world. However, one of the visible cues humans make and use to interpret each other's focus of attention is our eyes and gaze directions. Several psychophysical studies have even indicated that we are not blind to scenes which are outside of our focus of attention [16]. We are able to make simple judgements on objects to which we are not fixating on; objects which might only be visible in the outer corner of our eyes. So how do the eyes physically work? This chapter will focus on how we are able to see the world by researching the anatomy of the eyes and understanding the span of our field of vision.

4.1 Anatomy of the Visual System

Human eyes are quite distinctive and seem to have evolved in such a way that eye direction is particularly easy for our visual systems to perceive. The white *sclera* creates a high contrast to the iris and dark pupil, which is not seen often in other species [12].

The most basic features of the eye is displayed in Figure 4.1, where the eye is illustrated as if it was sliced horizontally with the nose underneath. The outside world is represented through light waves, which are processed in the eye and the brain. By simplifying the process, the overall flow consists of approximately 9 steps. These steps are highlighted in Figure 4.1 and Figure 4.2 which allows for a better understanding of the following descriptions of the individual steps.

(1) The light waves will first travel through the transparent and hard outer shell called the *cornea*

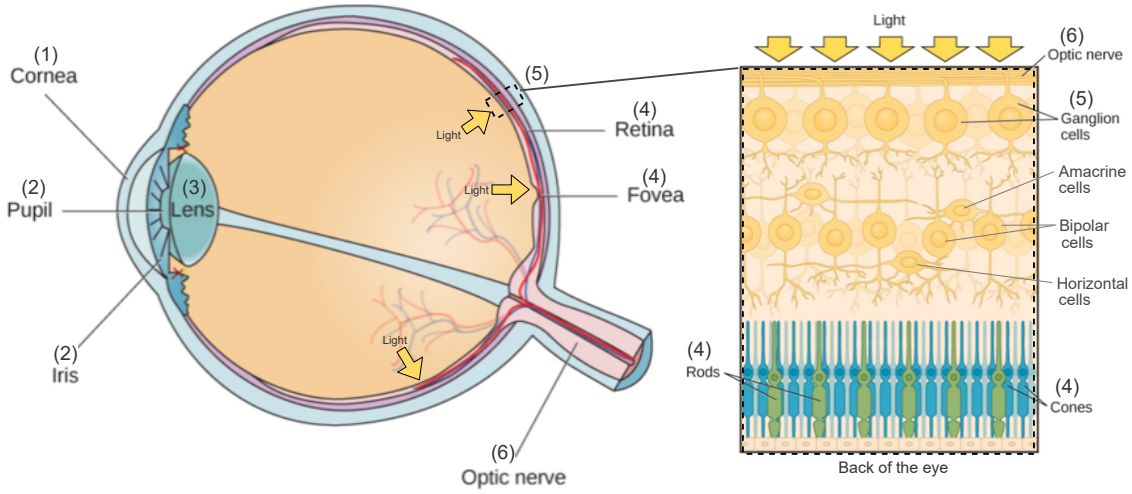


Figure 4.1: Images from [20] with small edits consisting of the numbers, the combination of the two figures, the light arrows and the depictions of amacrine, bipolar and horizontal cells.

[21]. The cornea is involved in focusing light waves and covers the front of the pupil and iris. The remaining part of the eye is covered by a white surface called the *sclera*. (2) After passing through the cornea, the light rays reach the lens through the pupil. The pupil's size is controlled by muscles connected to the iris, which is triggered by the light levels and emotional arousal [20]. The big eye movements are controlled by the brain, which is known as the oculomotor control, through 6 muscles on each eyeball [22]. These muscles allow the eyes to look in multiple directions even though the head is stationary.

(3) The *lens* is a curved, transparent structure that focuses and flips light rays so the image strikes the retina opposite of reality. The light rays from the left world view strike the right side of the retina and the top world view strike at the bottom. (4) The *retina* covers 180 degrees of the inner eye boundary and consists of a series of *photoreceptors* which are light detecting cells. Photoreceptors can either be *rods* or *cones*. Approximately 120 million rods are places throughout the retina and are triggered by low light levels, which enables humans to see in darker conditions. The cones are more sparse as the retina only has 6 million rods, which are mostly located in one section of the retina called the *fovea*. The cones are triggered by more light and are able to distinguish between colour. The high density of cones at the fovea enable people to see in more detail, so the eye moves until the focus of attention reaches the fovea. The density of the rods and cones are explained further in Section 4.2. (5) The photoreceptors send light signals to the brain after some initial processing. Figure 4.1 displays how the light is received by the cones and rods. The retina is turned inside-out, so the light actually passes over the optic nerve and ganglion cells to be caught by the photoreceptors. The photoreceptors send the signal back through a network of neural cells, where the most outer neural cells are called the ganglion cells towards the optic nerve.

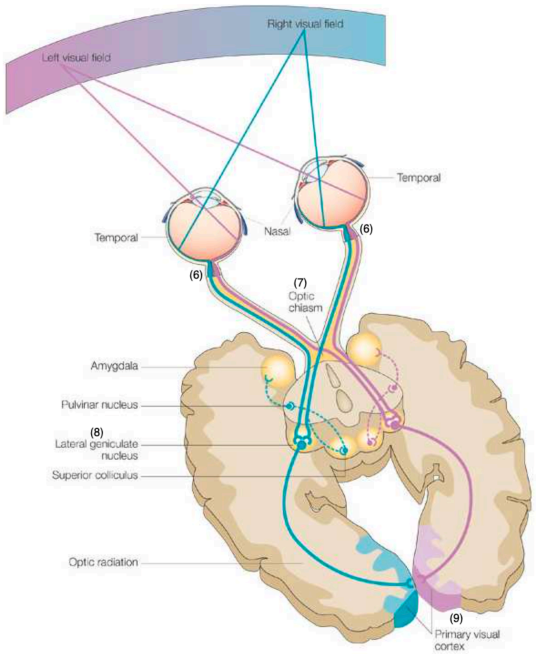


Figure 4.2: Image from [21] with small edits consisting of the numbers.

(6) The optic nerve is the pathway which transports the light information to the brain. The optic nerve runs along the inside of the eye, on the inside of the photoreceptors, and exits the eye through a spot called the blind spot. The blind spot does not have any photoreceptors, so our sight is actually not complete. However, given the fact that we have two eyes, where the blind spots do not overlap, the missing piece can be filled in by our brain. When only looking with one eye, the blind spot is filled in given the context, which is not always correct. (7) The optic nerve from both eyes will cross at the *optic chiasm*, where the light signal from the left sides and the right sides are exchanged. Figure 4.2 displays how the signal is passed on from the eyes to the brain. The left side of the brain analyses the visual information from the right side of the real world perspective from both eyes.

(8) The visual information from the optic chiasm is passed on to a part of the *thalamus* called the *lateral geniculate nucleus* (LGN), as seen on Figure 4.2 [21]. The LGN does not only recieve information from the eyes, but also the other senses. It serves as a router, by processing the information and sending it to the desired part of the brain. (9) The final processing step is in the primary visual cortex, which is located at the back of the brain. The primary visual cortex is connected with several areas of the brain and performs specialised functions. This is seen as the area of the brain where the individual visual perception of the world occurs.

4.1.1 Dorsal Stream and Ventral Stream

Visual processing is the ability to understand what the eyes see. Several abilities are defined as a part of the visual processing. Some of these abilities are mentioned below [22].

- **Visual closure**: The ability to know what an object is, even if only part of it is visible.
- **Visual Discrimination**: The ability to compare features and distinguish them from each other.
- **Visual Memory**: The ability to recall something seen recently.
- **Visual-spatial processing**: The ability to estimate how an object's location relate to yourself.

These abilities are only a segment of the abilities related to visual processing. So how can the brain interpret visual information from the eyes? As mentioned above, the processing of visual information happens in the primary visual cortex. The processed information exits the primary visual cortex and follows two major pathways in the brain [23]. The two streams bring meaning and understanding to the visual information, which is how humans create an inner perception of the world.

One of the streams is the *dorsal stream*, which is also known as the *where* or *how* pathway [22]. The dorsal stream controls the direction of attention and localisation of elements of interest. The dorsal stream is essential for coordination of bodily movements guided by the visual field [23].

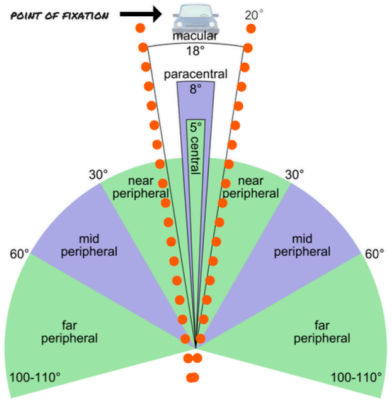
The other stream is the *ventral stream*, which is known as the *what* pathway [22]. The ventral stream is connected to the areas of the brain storing long-term memory (medial temporal lobe) and control of emotions (limbic system). The ventral stream therefore recognises and identifies visual elements. In addition, the ventral stream also judges the significance of the elements using the individual's memory.

The dorsal and ventral streams are bidirectional and interconnected allowing them to share information while processing in parallel [16]. The modelling of saliency can be compared to the streams, as the bottom-up salient stimuli is processed in the eyes' initial neural cells and the primary visual cortex, which extracts edges, color, shape etc. The top-down saliency is then processed by the ventral streams as it relates to memory and emotion. The processing of bottom-up salient stimuli will seem to pop out of a scene, which indicates that we are able to process a larger visual field than within the fovea. The speed of reaction with a bottom-up saliency-based form of attention is said to be between 25 to 50 ms per item [16]. In contrast, the speed of reacting to a task-dependent, top-down, form of attention is 200 ms or more per item, which is more than 5 times slower than the saliency-based form.

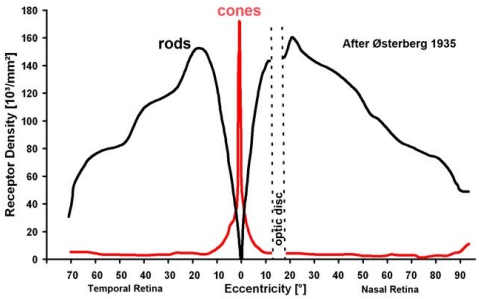
Vision is thereby not only dependent on the eyes, but consists of many parts including the combined functionality of the eyes, the eye muscles, and the brain [22]. The eyes are needed as the initial gateway for the light rays of the outside world. The muscles are essential to move the eyes and thus allow the focus to be in the desired direction, and lastly the brain is needed to understand, control and interpret the information.

4.2 Field of Vision

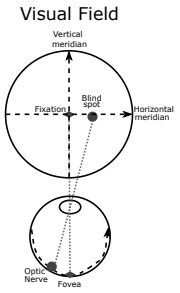
A child's ability to learn is highly dependant on their visual capability, as more than 80% of child's learning comes from vision [22]. The lack of vision can create movement, mobility, school and learning challenges in addition to the social and emotional difficulty that it brings. A person in the United States of America is deemed legally blind when they are unable to see more than 10 degrees from the point of fixation to each side [22]. A legally blind person is thus not always completely unable to see the world outside, but they are significantly less capable with a lower than average field of vision. The 20 degrees around the point of fixation is shown in Figure 4.3a, with 10 degrees on each side. Blindness is not only caused by damage to the eyes (Ocular level), but also by potential damage in the optic



(a) This shows the different fields of vision, where the red dots show the maximum boundaries for for being presumed legally blind. The image is from [22].



(b) This image is from [25], and shows the density of the photoreceptors along the horizontal meridian.



(c) This illustration is based on [26], but only includes the information relevant to this work.

Figure 4.3: These figures are from [22],[25] and based on [26]. They show the fields of vision, density of photoreceptors and horizontal meridian.

nerve transporting the information to the brain (Connection Level) and damage to the brain (Brain Level)[24].

The field of vision is commonly divided into 3 fields, as shown in Figure 4.3a:

1. **The Central (or Fovea) field** match up with the fovea in the eye, and is thereby the point where our vision is best. The field is about $2-5$ degrees and is defined as 100 acuity [22].
2. **The Paracentral (or Parafoveal) field** is the next area outside of the central field. It is often described as a number of letters' width around the central field. When reading, the paracentral field is typically about $2-4$ letters width to the left and up to 15 letters width to the right. It is typically within an area of 18 degrees.
3. **The Peripheral field** is the last field of vision. People are not able to see objects clearly within this field, however, it is essential for seeing movements. The peripheral field is divided into near, mid and far peripheral fields, as seen in Figure 4.3a. The peripheral field is an area about 110 degrees.

So, why is the acuity actually so different in the various fields? The reason for the diverse focus is the density of the cones and rods on the retina in our eyes. The density of the cones and rods along the horizontal meridian are shown in Figure 4.3b. The horizontal and vertical meridians are illustrated in Figure 4.3c, where the horizontal meridian crosses both the forvea and the blind spot. The majority of the cones are all at the fovea, which is marked in Figure 4.3b as 0 degrees. The cones are then drastically reduced and spread over the rest of the fields. The rods peak at 18 degrees from the center of the fovea, where there are approximately 160t000 rods/mm^2, and no rods are located at the very center of the fovea [25]. As the amount of cones and rods decline, the acuity of the vision thereby also declines. Thompson et al. studies the visual capabilities of humans in the peripheral vision, where they concluded that biological motion was perceived just as well in the peripheral field as in the central field [27]. So, when estimating what people are able to perceive of the outside world, it is essential not only to focus on the central field. Even though the central field is the focus of attention, the world in the paracentral and peripheral fields still influence our perceptions of the world.

4.2.1 Field of Vision Restrictions

The visual field is restricted by the anatomy of the head when trying to move the eyes in the desired direction. These restrictions can be divided into 4 portions, which are described below [24].

- **The Nasal Portion** is the area of the visual field towards the nose and is measured in degrees from the point of fixation. The area is normally limited to 60 degrees in the horizontal axis.
- **The Temporal Portion** is the opposite of the nasal, and thus the area towards the ear. The area is larger than the nasal portion and is normally limited to 100 degrees in the horizontal axis.
- **The Superior Portion** refers to the vertical space above the center of the visual field. This space is approximately 60 degrees in the vertical axis.
- **The Inferior Portion** is the final limitation and is the space underneath the center. This space is a bit larger than the superior portion with about 75 degrees in the vertical axis.

Figure 4.4 shows how we move our head and eyes in order to keep the attentive stimuli in the preferred viewing field.

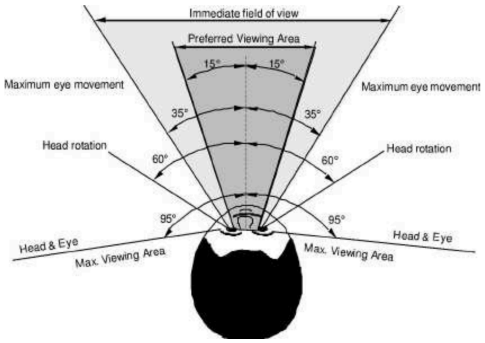


Figure 4.4: Image from [21] showing people's preferred movements when viewing a target.

The horizontal movement (yaw) is symmetric by allowing 35 degrees to the left or right, but the vertical field is not. The eyes are only able to pitch about 20 degrees upwards and 25 degrees downwards [21]. The variation in the vertical degree make people look slightly more downwards instead of completely straight.

These restriction do not allow both human eyes to see the same scene. The field of vision where both eyes are able to see the same scene is within the mid peripheral view, so approximately 120 degrees overlaps both eyes. The area seen by both eyes is also called the **binocular field**. The far peripheral field is thereby only seen by either the left or the right eye and is called the left or right **monocular field**. The binocular field allows people to better interpret the depth within the scene. When viewing an object from two different positions, the placement of the object changes for each viewpoint, which is called the *parallax*. The appearance of objects placed closer to the eyes will differ more than objects further away.

For this reason, humans' preferred horizontal viewing area is typically within the binocular field, as shown in Figure 4.4. People can comfortably perceive an area up to 30 degrees clearly, and will simply move the eyes within the immediate field of view which is 70 degrees. For areas larger than 70 degrees people will usually rotate their heads, and the maximum viewing area without moving our body is 190 degrees [21]. People are able to perceive a larger area within their peripheral view, but when looking at objects we will usually move our eyes accordingly.

Figure 4.4 is different from Figure 4.3a, as Figure 4.3a illustrates the different fields of vision when a person is fixating on an object. When fixating on an object, we are still able to see the area around the fixation, even though it might not be in high resolution. Figure 4.4 illustrates our preferred viewing area and how we will typically move our head and eyes to focus on different stimuli in a scene.

4.2.2 Kappa Angle

When a person is looking at an object, that object will line up with the fovea in the back of the eye to make the object clear. This line from the object of fixation to the fovea is called the *visual axis* [28].

However, it is actually not possible to observe this axis from an image of the eye, as the fovea is not completely lined up with the pupil. The angle that we are able to observe is called the *optical axis*, which is the direct line from the center of the pupil and to the back side of the eye (not the fovea). The fovea is often located about $4 - 5$ degrees horizontally and 1.5 degrees below the back of the eye [29]. The optical axis is found by estimating the direction a person's pupil is looking, which is made more simple by the contrast to the white sclera in the eye. The offset between the visual and optical axis is called the *kappa* angle, as illustrated in Figure 4.5.

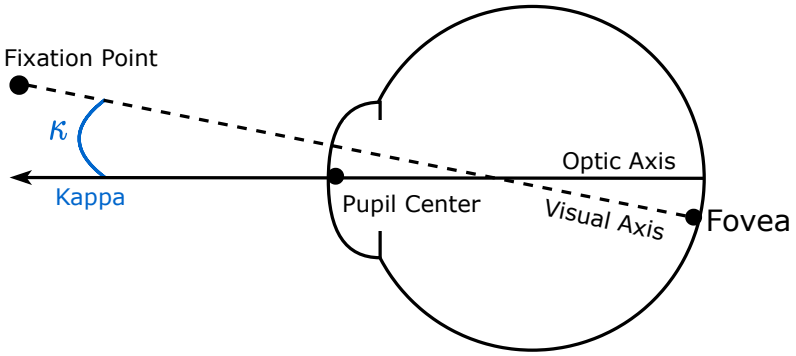


Figure 4.5: This illustration is based on [28] and [30], but only include the information relevant to this work.

The kappa angle is not the same for every person, and studies have shown that the angle for an average person can very up to 4 degrees [28]. As wearable eye trackers are currently able to estimate the gaze direction with about a 1 degree accuracy, a 4 degree variation has a large effect on the estimation. To compensate for this offset, eye trackers usually have an initial calibration, which estimates the individual kappa angle.

4.3 Summary

People's ability to percieve the world is dependent on the anatomy of the eyes, muscles and brain, which all contribute to our understanding of the outside world. The acuity within the field of vision is a product of the amount and types of photoreceptors, which are located on the back of the eye. The photoreseptors which are able to see colours are almost exclusively located at one point of the eye called the fovea. The resolution is at the highest when a stimuli is angled towards the fovea. Overall, the field of vision can be divided into three parts: the central, paracentral and peripheral fields, which all contribute with varying clarity. However, even though an action does not happen within the central field, studies have shown that movements are just as accurately perceived from the peripheral view, as the central view. So, when understanding what people are able to see, it is essential to include the near peripheral field of vision.

CHAPTER 5

Computer Vision

The previous chapters highlight how perception to some degree can be estimated based on gaze direction, which can be estimated using cameras or infrared sensors. Many gaze tracking models use basic computer vision techniques, particularly the fundamentals of deep learning, which are described in this chapter. It briefly mentions 5 different neural network architectures, which are used within the gaze tracking methods mentioned in Chapter 6.

5.1 Deep Learning

Deep learning has changed the way data is viewed by teaching machines to outperform humans in some tasks [31]. Artificial Intelligence is the overarching field containing machine learning and, in turn, deep learning. Deep learning has over the last two decades proven how powerful it is at solving tasks in computer vision, speech recognition, medical information processing, natural language processing (NLP), cybersecurity, and many more [31].

Briefly explained, deep learning is a technique calculating weights in a neural network based on data. A neural network is a network of small perceptrons that receive inputs from external sources, adds weights, biases, and a nonlinear function before returning an output. These perceptrons are divided into multiple layers where the initial layer receives the input, followed by a number of hidden layers, and finally the output layer. A network is considered deep when it contains more than one hidden layer.

The weights and biases of each perceptron are trained by using stochastic gradient descent. Stochastic gradient descent calculates the loss of the final result and uses it for back-propagation, moving backwards through the network and updating all the internal parameters. The deep neural network can have multiple different structures depending on e.g. the loss function, the amount of hidden layers, the amount of perceptrons, the nonlinear functions etc. Some of these structures have proven quite useful within computer vision, and has in turn been used within gaze tracking.

5.2 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a specific kind of neural network that utilizes the grid-like layout of data such as images. The modern CNN was first introduced in 1998 by Yann LeCun et al. [32], and after impressive results of AlexNet in 2012 [33], the technique revolutionized image classification. The structure of CNN consists of three sections; convolutional layers, a fully connected layer, and often a softmax function.

The *convolutional layers* slowly extracts features from the data to gradually create a more advanced and complex input. Convolutions use kernels, also called filters, of weights which slide over the image with a specific stride. At every stride, a matrix multiplication is performed and summed up, resulting in a single number being added to the feature map. When the kernel has slid over the entire image, the result is a feature map, which emphasizes different areas of the image. The dimensions of the feature map is dependent on the size of the kernel, the size of the step a stride performs, and whether a padding is added to an image. An activation function is used on the feature map to create a non-linearity in

the network. Finally, a convolutional layer is mostly followed by a max pooling, which reduces the dimensions of the feature maps by selecting the maximum number within a small sliding window.

Multiple convolutional layers can by stacked and followed by *fully connected layers*, also known as the output layers, which interprets the feature map. This will classify the input dependent on the extracted features from the convolutions. A normal fully connected layer cannot receive n-dimensional data, so the feature maps have to be flattened into 1-dimensional data first. The output, in a classification task, will mostly be assigned a probability for every possible classification through a *softmax function*. The softmax function basically ensures that the probability classifications are between $[0-1]$ and adds up to 1.

The trends for CNN in early 2010s was to make the networks deeper to improve the performance [31]. However, deep networks are notoriously hard to train due to the vanishing gradient problem. The vanishing gradient problem occurs when the gradient is back-propagated to earlier layers, and the repeated multiplication can make the gradient infinitely small and thus vanish. In 2015 Kaiming He developed a new structure called residual network (ResNet), which is designed to be more stable for training deep networks [31].

5.3 Residual Network (ResNet)

Residual networks (ResNet) are designed to avoid vanishing gradients by creating a traditional feed forward network with an additional connection called the residual connection. A ResNet consists of several basic residual blocks, which in turn contain different convolutions and activation functions depending on the architecture of residual networks. These blocks' output is typically the sum of the output from convolutional layers and the original input. This gives the algorithm an opportunity to be selective on when to utilise specific blocks, and thereby have a smaller chance of encountering the vanishing gradient problem in gradient descent. In 2015 ResNet was the deepest neural network architecture seen yet and outperformed all previous state-of-the-art solutions [31].

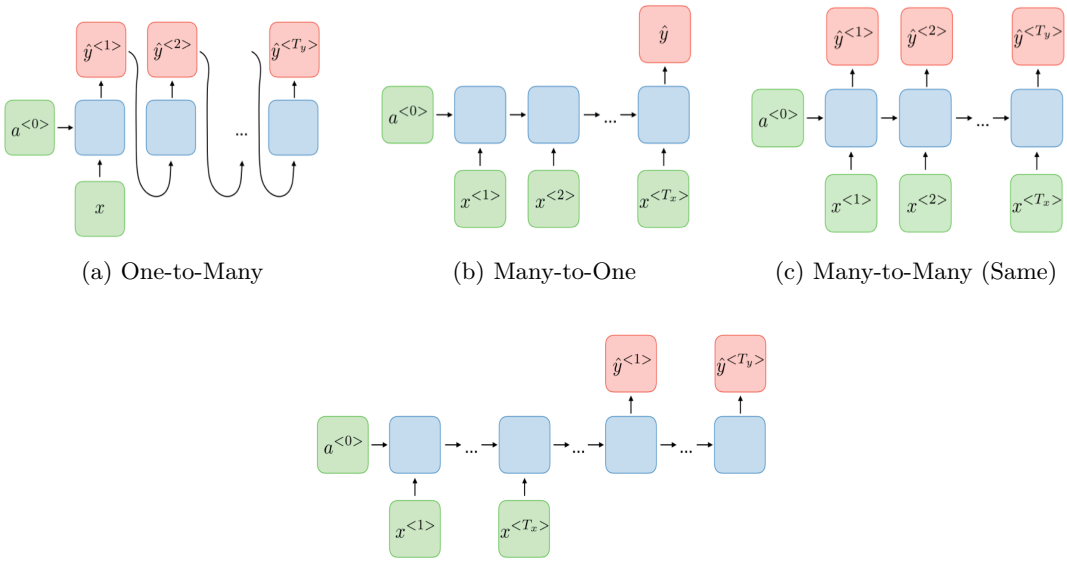
5.4 Recurrent Neural Networks (RNN)

The sequential order of the data does not influence a typical feed-forward neural network. However, it can occur that the order of individual observations contributes to the interpretation. Recurrent Neural Networks are designed to share information from previously hidden layers and pass it on to the following time-steps. Passing data from previous time-steps allows the model to couple the data observations together and calculate a more informed output.

The architecture of RNN allows for structures that vary in the number of input and output values, which has not been seen before. In an original feed-forward network, the structures are simply one-to-one e.g. give one image, receive one word [34]. Examples of RNN applications show four reoccurring architectural structures of RNN: one-to-many, many-to-one, many-to-many (different), and many-to-many (same) are shown in Figure 5.1 [35].

An application of a *one-to-many* architecture is seen in melody and arrangement generation [36]. In this particular music generation task, a one-to-many sequences generation (OMSG) task is created. The authors of the paper wanted to generate music with correlation between tracks (drum, bass, guitar etc.) which they did by using one melody as input and generating several tracks as output. Thus, a one-to-many application of RNN.

The second application is *many-to-one*, where an example can be found within sentiment analysis [37]. The sentiment analysis in this example can have the input at either character or word level. At character level, the input was a list of characters in a sentence, while word level is a list of words in a sentence. The output of the sentiment analysis would calculate whether the sentence was positive or negative, thus making this sentiment analysis the type many-to-one.



(a) One-to-Many (b) Many-to-One (c) Many-to-Many (Same)

(d) Many-to-Many (Different)

Figure 5.1: These figures show four reoccurring architectural structures of RNN. The illustrations are from [35].

The third application is *many-to-many (same)*, which is seen in entity recognition [38]. Entity recognition is the ability to classify data according to the context (e.g. recognising people, cars, etc. in the data). An example is the sentence *I love Vietnam*, where each individual word can be translated to *O O B-Loc*. *O* translates to no entity and *B-Loc* translates to belonging to the entity *location* [39]. Every word in the sentence is translated to an entity, and thus a many-to-many (same) RNN.

Finally, the *many-to-many (different)* application is found within machine translation [40]. When translating context, the number of words are often not the same in different languages. This tasks is possible given RNN with an architecture of many-to-many (different), as it allows for the input and output to vary in size.

The equation for calculating the individual nodes' output in a traditional RNN can be seen in Equation 5.1. The output for the given time-step (h_t) is the result of a function (f_W) given the perceptron's weights (W) with two inputs. In the Vanilla version of RNN, see Equation 5.2, an activation function, tanh, is used on the weights multiplied with the hidden state from the previous time-step (h_{t-1}) and the input (x_t).

$$h_t = f_W\left(h_{t-1}, x_t\right) \tag{5.1}$$

$$h_t = \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t\right) \tag{5.2}$$

Like with CNN, RNN also has problems with vanishing and also exploding gradients, where the gradients become infinitely large. Gradient clipping is a method that simply ensures that the gradient cannot become larger than a set value. This can remove the chance of exploding gradients even though it might not be the most elegant solution. To avoid vanishing gradients one can implement another network called Long Short Term Memory (LSTM).

5.4.1 Long Short Term memory (LSTM)

Long Short-term memory (LSTM) is another model designed to avoid the vanishing gradient problem. The key concept of LSTM is the introduced cell state. The cell state consists of four gates which are either added or removed. Equation 5.5 explains how the gates are used to calculate the cell state and thus the result in the new output.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} (W_{hh} h_{t-1} + W_{xh} x_t) \tag{5.3}$$

$$c_t = f \odot c_{t-1} + i \odot g \tag{5.4}$$

$$h_t = o \odot \tanh(c_t) \tag{5.5}$$

The gates are typically called input gate (i), forget gate (f), output gate (o) and the final one was dubbed gate gate (g) by Justin Johnson[34]. This model avoids the vanishing or exploding gradient problem, as the back-progagation will go through the cell state where the weight can be updated given elementwise multiplication through the forget gate (f). As the forget gate varies from each time-step it will not have the continual matrix multiplication with identical values in each time-step seen in the vanilla RNN (Equation 5.1).

5.4.2 Recurrent Convolutional Neural Network (RCNN)

RNN and CNN have been combined to Recurrent Convolutional Neural Network (RCNN) in 2015, which advanced the state-of-the-art accuracies of object recognition [41]. The recurrency in CNN allows the algorithm to boost the performance by enabling units to be modulated by other units in the same layer. The key feature in RCNN is the recurrent convolutional layer (RCL). RCL creates recurrent connections within every convolutional layer of the feed-forward CNN while still keeping the number of parameters constant when using weight sharing between layers like RNN. The results from this network demonstrates the advantage of the recurrent structure over purely feed-forward structures for tasks such as object recognition [41].

5.5 Summary

Computer Vision is not confined to deep learning techniques and can extract features within an image without neural networks. However, the best results are created based on neural networks, which are utilised for gaze tracking, which will be discussed in Section 6.

CHAPTER 6

Gaze Tracking

6.1 Gaze Tracking Terminology and Taxonomy

A survey from May 2020 by Dario Cazzato et al. on computer vision contributions for human gaze estimation and tracking describes the techniques used in various research papers dating back from 1792 to 2020 [42]. The survey describes the lack of a standardized terminology and thereby clarifies terms used interchangeably such as *eye detection, eye localization, eye tracking, gaze estimation, gaze tracking*, and *point of regard*. The terms used in this paper are primarily *eye localization*, *gaze estimation*, *gaze tracking* and *point of regard*. *Eye localization* is the technique to detect and locate individuals' eyes in an image and create bounding boxes around the eyes. This is typically used as a preliminary step, as locating the eyes is necessary to estimate what the individual is looking at, which is called *gaze estimation*. This is often defined as a 3-dimensional vector from the location of the eyes and toward the direction the gaze is looking in. When the gaze is estimated continuously over time, the act is *gaze tracking*. Finally, the *point of regard* is the object or area the individual is looking at.

6.1.1 Gaze Tracking Taxonomy

In addition to the terminology, the survey also suggests a general taxonomy to standardize the different aspects of gaze tracking.

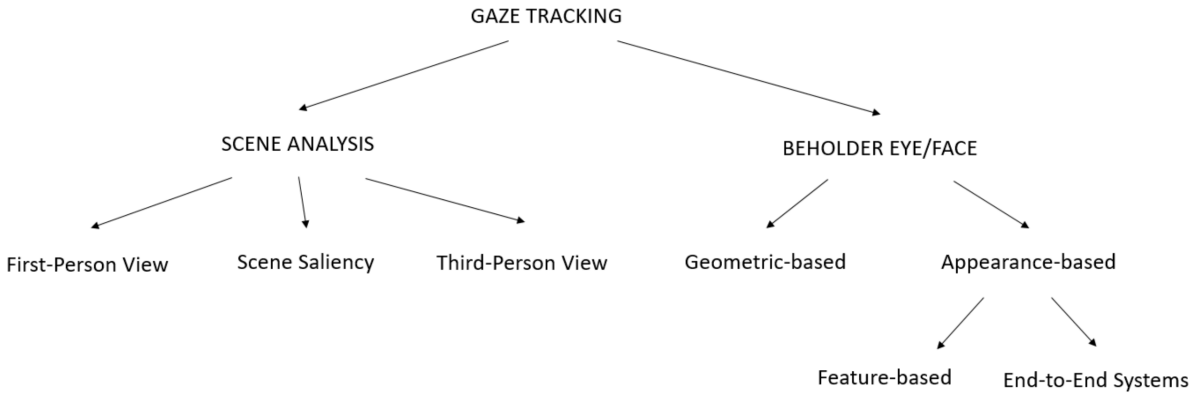


Figure 6.1: Suggested Gaze Estimation Taxonomy from [42].

The schema, seen in Figure 6.1, categorizes the gaze tracking technologies into two categories; scene analysis and beholder eye/face. Scene analysis describes when the data from the acquisition devices displays the scene the beholder is viewing. This scene can contain the targeted beholder or simply be the scene the individual sees. The second category is when the received data shows the eyes or face of the beholder. This will thereby not contain the scene the individual is looking at, but the person beholding the scene.

An overview of the different categories and sub-categories are illustrated in Figure 6.2. The model

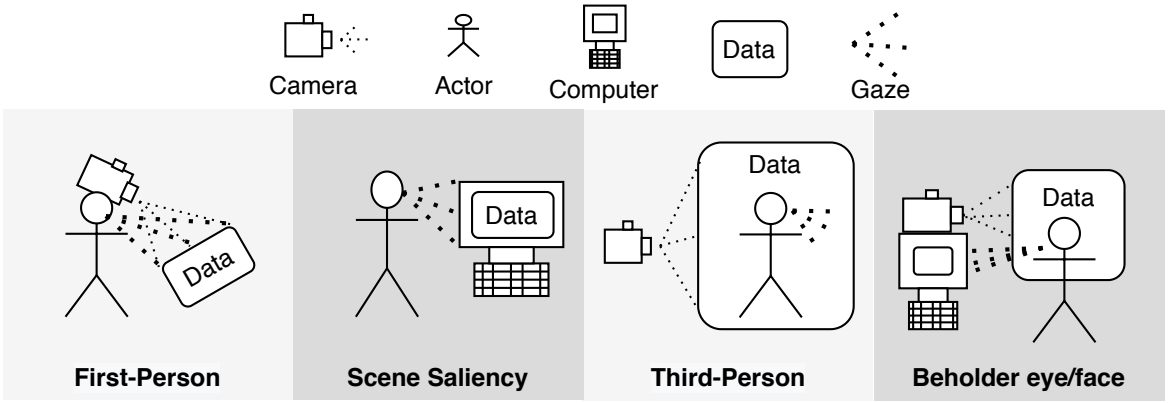


Figure 6.2: This overview was created based on the suggested gaze estimation taxonomy types from [42]. The model shows the three sub-categories within scene analysis (First-Person, Scene Saliency, and Third-Person) as well as the setup for beholder eye/face. Beholder eye/face's sub-categories all have same setup.

shows how the sub-categories relate and differentiate to each other, which are described using examples in Section 6.2 and Section 6.3.

6.2 Scene Analysis

Scene analysis consists of three groups; First-person view, scene saliency and third-person view. Each of these groups will be described with brief examples of existing models, datasets, and an analysis on how they relate to this study.

6.2.1 First-Person View

For *first-person view* the camera is often placed on the beholder, e.i. a head-mounted piece, so the image from the device is equivalent to the beholder's view. An example of the setup can be seen in Figure 6.2, where the data consists of images of the scene viewed by the beholder. The egocentric vision is often used to locate task-relevant foreground objects the beholder is manipulating. The gaze tracking will mostly rely on eye, head and hand coordination to estimate where the beholder's attention is focused. The study within first-person view is fairly new as the popularity and technology of wearable devices is gradually increasing. It is common for technologies within first-person view to predict gaze using framed objects (objects surrounded by bounding boxes) [42].

A paper published in 2019 [43] uses the first-person view approach, where an example of the data can be seen in Figure 6.3. The suggested network is divided into two parts: Object recognition and gaze estimation. The object recognition part uses a pre-trained network called YOLO9000 [44] to extract features from the input data. The gaze estimation part has two levels, a coarse and a fine-grained level. The intent of the coarse grid location is to estimate the location of the gaze at an object level, while the fine-grained gaze should predict the probable and precise gaze coordinates. It was discovered during training that combining the losses from the two levels improved the performance of the overall network.

The coarse level, object level (classification), is estimated through a deep learning network using several convolutional layers, specifically deformable convolutions, and a softmax function. Convolutinoal layers are described in Section 5.2, whereas deformable convolutions are quite similar. The main difference is that deformable convolutions adds an offset to the original algorithm [45]. This offset allows for better results when recognizing objects in the data that might be augmented or transformed

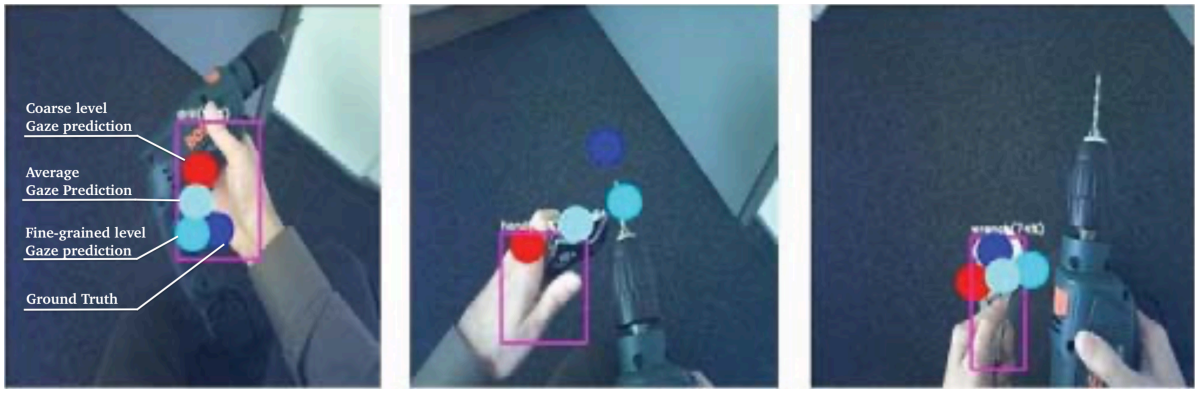


Figure 6.3: Scene Analysis - First-person view example from [43] with the captions of the dot added. The red dot is the gaze predicted by the coarse level, the middle/regular blue color is by the fine-grained level, the light blue is the joint prediction, and finally the dark blue is the ground-truth.

due to a specific angle. The purpose of the softmax function at the end of the network is to gain a probability distribution of the gaze on the objects as the final output. In contrast, the fine-grained level estimates the probable gaze coordinates (regression). Like the coarse level, the fine-grained level uses two deformable convolutions. In addition, the fine-grained level also uses two fully-connected layers to create the coordinate output. The result of the network can be seen in Figure 6.3, where a couple of frames shows the two coordinates of the coarse and fine-grained outputs as well as the averaged final prediction.

The authors of the paper [43] used three different datasets for training the network, where two are publically available and the third was created by themselves. They used the *GTEA-plus Dataset* [46] which is composed of 35 videos using 5 different subjects. The subjects perform meal preparation activities, which is annotated by using SMI eye-tracking glasses as well as hand annotations of the actions in the frames. The second dataset is the *Object Search (OS) Dataset* [47] containing 57 videos of object search and retrieval tasks by 55 different subjects. The data is annotated using eye trackers, which provides the ground truth of the point of regard. The final dataset, which was created by the authors, is called the *Tools Search Dataset*. This dataset is a collection of 4 videos, each about $3 - 4$ minutes in length, recorded using an SMI eye-tracker. These videos are in an industrial setting of finding tools and fixing problems in the drill machine, as seen in Figure 6.3 [43].

The performance of the network is evaluated using an Average Angular Error (AAE) metric, which calculates the angular distance between the ground-truth and the prediction. The AAE for the GTEA-plus Dataset was 6.2 degrees, whereas the previously reported best accuracy was 4.0 degrees. However, for the OS Dataset the model achieved state-of-the-art performance with a AAE of 10.2 degrees, compared to the previous best of 10.6 degrees. In addition, the network is reported to be able to run in real-time performance (13-18 FPS) [43].

The first-person view is interesting and could potentially be useful in many scenarios where individuals are working with objects. However, this study is based on a robot's point of view to predict peoples' field of vision. Using the techniques from first-person view would require everyone interacting with the robot to use a head-mounted camera that sends data directly to the robot. This is not ideal in a real life scenario as the intention of the robot is to help people, not create an additional inconvenience. A more distinct option is to use the sensors already provided through the robot as the data.

6.2.2 Scene Saliency

The second sub-category is *scene saliency* that analyses an image (or screen) viewed by the beholder [42]. Figure 6.2 illustrates this scenario, where the beholder is shown an image and the algorithm predicts the

areas of the image a person would initially fixate on. Unlike first-person view, the data is not retrieved from a head-mounted camera, which will change the view based on head movements. So the concept of eye, head and hand coordination cannot be used within scene saliency. Scene saliency has been studied extensively in literature over the years and has multiple applications including advertising, interface optimization, video and sound surveillance, medical imaging, etc. [23].

Overall, there are three techniques that are used to determine a saliency map; bottom-up, top-down, or a hybrid approach. The bottom-up approach predicts eye fixation areas by extracting the low-level features (colour, intensity, orientation etc.) in the image. This is quite similar to how human attention reflectively focuses on low-level features. In turn, the top-down approach locates high-level features (objects, humans etc) within the frame. As mentioned in Chapter 3, this top-down approach is not based on memory or emotion as the human brain will interpret it. Finally, the hybrid combines both top-down and bottom-up approaches to improve the performance of the individual.

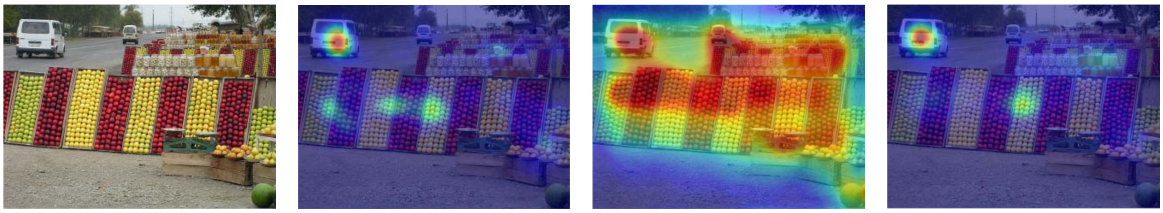
A comparison of top-down (high-level features) vs. bottom-up (low-level features) have been explored in 2017 [48]. The study created two different neural networks: one was trained using a pretrained object recognition network for detecting high-level features (DeepGaze II), while the other was trained to focus on contrasts in the images (ICF). The result in this study showed how DeepGaze II performed extremely well with, at that time, state-of-the-art results, while the ICF was comparable, but not as impressive. Upon further investigation, the authors found that the ICF performed better on 10% of the data, specifically, on the images not containing objects or containing high contrast distractions. Their conclusion was that high-level networks could be improved by training high and low-level models jointly [48].

Figure 6.4: Scene Analysis - Saliency view example from [49]. The images from the left to the right shows: the original image, the ground-truth of the eye fixations, the ground-truth combined with the saliency map, and finally the suggested saliency map respectively.

An example of a hybrid approach to scene saliency is shown in Figure 6.4, which shows the original image, the ground truth and the predicted saliency heatmap [49]. The example emphasises how peoples' attention is drawn to the little girl's face and the object in her hand. The model is designed using an end-to-end system which combines several loss functions to gain the most accurate prediction. It uses an attentive convolutional long short-term memory network (Attentive ConvLSTM) [49]. Attentive ConvLSTM enables the algorithm to iteratively focus on relevant spatial locations to refine saliency features. In addition, the model has a learned prior that models peoples' bias, which are focusing their attention on the center of an image. The strength of this model is the versatility of the types of images it can analyse. The report shows how this model performs well for both images containing objects (high-level features) and landscapes (low-level features).

Deep neural networks are very dependent on the datasets. Four of the most popular datasets within saliency are *SALICON*, *MIT1003*, *MIT300*, and *CAT2000* [49]. SALICON is the largest of the four datasets containing $10,000$ training images, $5,000$ validation images and $5,000$ testing images, while the smallest of the four datasets is MIT300 containing 300 natural images. MIT1003 and MIT300 has annotated the data using eye-tracking techniques from respectively 15 and 39 observers. In comparison, SALICON has the data annotated by simulated mouse movements [49]. The final dataset CAT2000 contains a large variety of categories such as Cartoons, Art, Satellite, Low resolution images, Indoor, Outdoor, Line drawings, etc. [49].

A comparison of the two mentioned networks can be seen in Figure 6.5. DeepGaze II maintains state-of-the-art results when evaluating Area under the ROC curve (AUC) for the SALICON test set at 0.885, while the attentive ConvLSTM's result for AUC is 0.883 [49]. However, for other evaluation matrics, the attentive ConvLSTM outperforms DeepGaze II and numerous other networks. AUC do not penalize low-valued false positives giving a high score for high-valued predictions placed at fixated locations while ignoring the others. An evaluation matrics, which is sensitive towards false positive and false negatives, is Normalized Scanpath Saliency (NSS). For the NSS evaluation the attentive ConvLSTM preforms significantly better with a result of 3.204 compared with DeepGaze II with 1.336 [49].



(a) Original image from the SALICON validation dataset.

(b) The ground truth.

(c) Results from DeepGaze II.

(d) Results from the attentive ConvLSTM.

Figure 6.5: Comparison of the results from two state-of-the-art networks predicting saliency maps [49].

Another interesting approach to determining the saliency map is focused on the human emotion-evoking objects in the images, as these tend to draw our attention [50]. The approach seems too subjective, as individuals have different emotions towards objects, which could thereby cause drifts on saliency outcomes. This solution simply shows how different approaches can be when solving the same problem.

Overall saliency maps are very capable, as they are able to highlight areas our attention is drawn towards. Saliency maps have been used to enable gaze tracking as a classification task between the highlighted saliency areas in an image, and thus only use the gaze to determine which area is the most probable for the gaze. In order to determine whether an object is within a person's field of vision a classification task could be used. However, based on the knowledge of cognitive load, it is possible for people to perceive more than one object at the time, and thus not seek a binary answer.

6.2.3 Third-person view

Finally, the last category within scene analysis is *third-person view*. Third-person view is where both the beholder and the scene viewed by the beholder is in the image. Unlike scene saliency the model does not predict what a person viewing the image will look at, but what the individual **in** the image is looking at [42]. This is demonstrated in Figure 6.2, where the intention is to predict the gaze directions of the people in the images.

There are multiple techniques to predict the individuals' gaze direction in images. One method combines the direction of the gaze with a saliency map. The saliency map is used to divide the image into areas, which in turn can create a classification problem by detecting which saliency area a person is looking at. This is to utilize existing technology to simplify the task at hand [42].

An interesting example of third-person view gaze tracking is a study by Sümer et al. [51]. The intention of the study is to determine the joint attention between humans, which can be seen in Figure 6.6. The model is designed in an end-to-end fashion that takes a raw image as input and generates a two-channel likelihood distribution (Attention Flow). The first channel detects and locates the faces of people in the image, while the second determines the joint attention of the people, if there are any. The model is designed using three modules; encoder, attention flow generator and saliency-based ground

truth generation. The encoder extracts faces and objects in the scene through object classification tasks. The attention flow generator estimates the likelihood of joint attention with the encoded images as input. Finally, the saliency-based ground truth generation use saliency maps, specifically DeepGaze II mentioned in Section 6.2.2, to produce Pseudo-attention maps to emulate the ground truth. By employing saliency estimation in the method, they leverage information of the relative important regions, which can also be salient for the persons in the scene.

Figure 6.6: Scene Analysis - Third-person view example [51]. The Image to the left is the original image with the ground truth empathized by the green bounding box, and the input to the model. The two images to the right are the outputs, first the face likelihood and then the joint attention map.

The model is trained and evaluated using various different datasets. They discovered that the encoder performed better when they used transfer learning from an object classification trained on the well-known dataset ImageNet. For evaluation they used the Video Co-Attention (VideoCoAtt) dataset, which contains 380 RGB video sequences from 20 different TV shows [52]. The dataset has been annotated manually to determine the joint attention areas within the images. The joint attention network proved to outperform previous models with a prediction accuracy of 78.1% compared to the previous state-of-the-art, which had an accuracy of 71.4%.

Third-person view scene analysis is the most important category given this particular case, as the intention is to estimate what people (who are in the images) are looking at. The setup, having a third-person camera view, is identical to the setup available by the robot. Thus, utilizing methods within third-person view to broaden the intention from *What is the people looking at* to *How likely is it that people can see this object* seems like a feasible task. Additional models within this sub-category are described in Section 7.1 in more detail, as they have been implemented in an experiment described in Chapter 7.

6.3 Beholder Eye/Face

The second main category for estimating gaze tracking, as shown in Figure 6.1, is the beholder eye/face. Though this category has several sub-categories, the overall setup is similar for all. This setup can be seen in Figure 6.2, where the camera is facing the person, while placed in the direction of the item the person is fixating their gaze upon. The beholder of eye/face category, which will be described in the following sections, is thus the opposite of first-person view. So instead of having a camera placed on the person, the camera is located on / close to the device being viewed by the person and the data is thereby an image of the person.

The techniques use to interpret the data within this category can either be calculated using the physical geometric of the eye, or the appearances of the face in the images. Gaze tracking that focuses on individuals' faces have previously been done using intrusive methods (e.g. electrooculogram) in addition to the non-intrusive methods (e.g. infrared illuminations etc.). The intrusive methods are not used as much today, as the non-intrusive methods are very precise [42]. Some of the non-intrusive methods, which are using cameras to collect data, rely on computer vision techniques to extract information from the images similar to the previously mentioned categories for gaze-tracking. Section 6.3.1 and Section

6.3.2 describe examples from these non-intrusive methods for both the geometric approach and the appearance based approach.

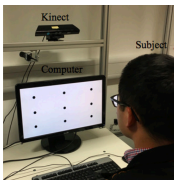
For most part, techniques within the beholder eye / face category relies on the face of the beholder being clear and often within a certain distance of the camera. In addition, the image will be directly facing the person of interest, so the scene the beholder is viewing is not visible within the image. The intention with beholder eye/face is thus to locate, with an accurate precision, where the beholder's main focus of attention is e.g. on a computer screen.

The goal of this thesis is to allow a robot to estimate whether people are able to see certain objects within the scene. A camera will therefore face both the people and the scene from a third-person view. The data available will vary in conditions and quality as the people will stand at different distances and angles to the cameras. In addition, the angle of the camera will not always be located at the object of interest, as the object could be placed anywhere within the scene of the image.

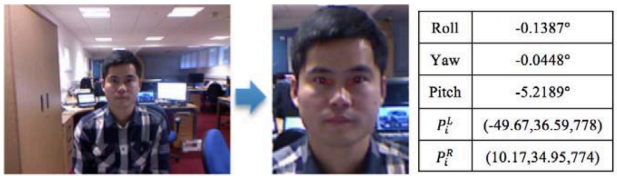
The beholder eye / face category could be suitable if the intention was to discover where on the robot itself people were looking, e.g. on the touch screen on the robot's chest. This would ensure that the person is close enough to reach the touch screen and thereby allow a more ideal scenario to use beholder eye/face techniques. Even though this described use is not the focus of this work, the category is a critical part of understanding gaze tracking. The two subcategories will therefore be described through examples in the following sections.

6.3.1 Geometric Based

The *geometric based* approach estimates the gaze direction through geometric reasoning of 3D models of the head and/or eyes [42]. Computer vision is used to detect the eye as well as the head position, eye position, pupil center, iris edges etc. In some cases, data is manually inserted for calibration in the initial stages. This data is used to personalize the estimations, as individual peoples' gaze direction vary based on the individuals' eye shape [29]. Overall, the estimation often calculates the point of regard on a scene opposite of the beholder using equations based on the known geometric distances in the scene as 3D models.



(a) The setup for the calibration of the algorithm.



(b) The input data and the values being calibrated.

Figure 6.7: Beholder Eye/Face - Geometric based example from [53]. The figure shows how the method is calibrated based on the input data.

An example of a model designed on a geometric method is called *Two-Eye Model-Based Gaze Estimation from a Kinect Sensor* by Zhou et al. [53]. Figure 6.7a shows the setup for this experiment. The model estimates the gaze of both eyes individually, and combines the results by taking the average at the end. Initially, the model is calibrated to the individual person by having the person look at a limited about of dots on a computer screen. This is illustrated in Figure 6.7b. The calibration is used to model the person's facial features, so the model can calculate the direction of the gaze given small head-movements. The five parameters, roll, yaw, pitch, P_i^L, and P_i^R, are estimated and used to calculate the eyeball centers ($O_e^{L/R,H}$) in the head coordinate system and Kappa angles ($\alpha^{L/R}$,$\beta^{L/R}$). The Kappa angles are different for the two eyes, and is the angle between the angle the actual direction

the individual is looking, and the direction calculated based on the image features. Thus, all the above parameters can then be translated to a location on the screen the beholder is estimated to look at.

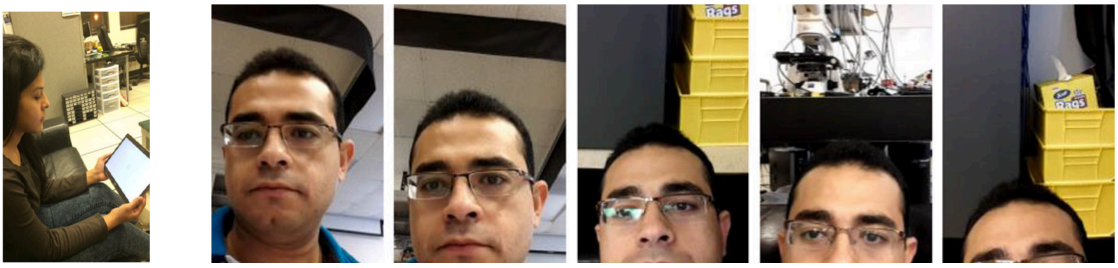
The data is based on an kinetic sensor, which have a range between 80cm to 400cm from the sensor to the person, thus the individual does not exceed this limit in the experiment. The model allows for limited head movements to maintain a high accuracy, as too large head movements will provide a smaller view of both eyes. Overall, the model performs with an average accuracy of 1.99 degrees over 11 test subjects. In comparison, the second best model mentioned in the paper had an average accuracy of 1.4 to 2.7 degrees. As the data is based on 3D geometric calculations, a dataset has not been necessary to train a neural network. Geometric gaze tracking can thus be calculated without the use of deep learning, whereas this is the central technique for most of the other tracking methods.

6.3.2 Appearance Based

The other approach within beholder eye/face is *Appearance Based*, which estimates a direct mapping between the eyes/face and the point of regard using computer vision techniques and not specific geometric models. This approach is divided into feature based systems and end-to-end systems as seen in Figure 6.1. The main difference between the two systems is whether features from the images are extracted prior to the gaze tracking algorithm. The following subsections, Section 6.3.2.1 and Section 6.3.2.2, describe examples within each system.

6.3.2.1 Feature Based Systems

It has been shown that low-level gradient features are well-suited to capture variations in the appearance of the eye, as mentioned in Section 4.1.1. These low-level features are utilized when creating *feature based* models. The feature based technique is used in an example by Huang et al. that estimates the gaze of subjects looking at a tablet [54]. The input for the model is images from the built-in camera within the tablet, which creates a simple setup.



(a) The setup when people participate in the experiment.

(b) Cropped examples of the data, which show the diversity in the visibility of the face and the potential reflections in from the glasses.

Figure 6.8: Beholder Eye/Face - Appearance Based, feature-based example from [54]. The figures show the setup and examples of the data inputs.

The setup and examples of the data are shown in Figure 6.8. Unlike the example mentioned in Section 6.3.1, the model does not need calibration and the study has no restraints on the individual's amount of head movement. This model is built in three parts; prepossessing, feature extraction and regression estimation. The prepossessing step locates the beholder's eyes, and crops the area of the eyes with a small margin to the size of 30×100 pixels. The feature extraction step then locates the image's features through *multi-level Histograms of Oriented Gradients* (mHoG) and a dimensionality reduction. Multi-level Histograms of Oriented Gradients is a machine learning technique for extracting a feature

vector describing an image. It calculates the gradient magnitude and the direction the magnitude is changing in. These two features combined will result in a feature vector and thereby mHoG. The gradient magnitude is calculated through the horizontal and vertical distribution of intensity within the image. This can thereby locate edges of objects within the image, if there is a high contrast.

The final step is a regression estimation. Four regression models were tested, where the best results where gain through a Random Forest regression model, where two separate regressors are trained to either estimate the horizontal coordinate or vertical coordinates. The combination of these regressors produces the final estimation of the point of regard.

To train the deep learning network, a dataset was constructed dubbed TabletGaze. TabletGaze contains 51 subjects, each with 4 different postures and 35 gaze locations. The subjects vary in race, gender and in their need for prescription glasses, which all diversified the dataset. An example of the data is seen in Figure 6.8b, which demonstrates how the eyes of an individual are not always visible from the tablet's camera given the user's position. The algorithm will thereby be trained to identify the point of regard even when the eyes are not visible, or when the glasses' reflection obscure the view of the eyes.

The paper evaluated different feature extractors and regression models to find the best solution. Overall, the results where greatly influenced by how much of the face was visible within the images. Given the TabletGaze dataset the best result, using euclidean distance as the error metric, was a mean error of 3.17 cm given 100.000 images. In addition, the paper evaluated on the performance given glasses, race and posture. They concluded that wearing glasses had an additional mean error of 0.4 cm, in comparison to people without glasses. The result for race showed similar results as the results for Asian compared to Caucasian had a noticeable difference of approximately 0.3 cm. Finally, the body posture of the subjects had quite equal results, with the laying position having the highest variation with a mean error of approximately 0.1 cm. They believe the reason the laying position varies is the individual's method of holding the tablet.

6.3.2.2 End-to-end systems

The other appearance based technique is creating an *end-to-end* system, so simply having the image as input without specific feature extraction. An example of such a system developed by Palmero et al. published in 2018, with the model illustrated in Figure 6.9.

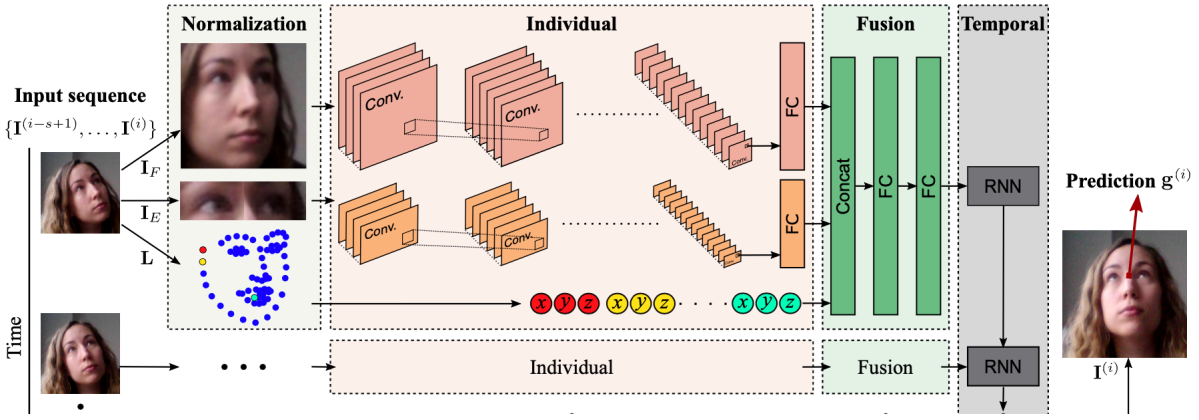


Figure 6.9: Beholder Eye/Face - Appearance Based, end-to-end example from [55]. The input data is shown to the left and the architecture of the network is illustrated to the right.

The model by Palmero et al. is based of the EYEDIAP dataset, which they used for training and testing of the model [55]. The EYEDIAP dataset is a publicly-available dataset which contains video sequences with a wide range of head poses and showing the full face. It consists of 3-minute videos

of 16 subjects who look at two types of targets: a continuous screen target on a fixed monitor, or a floating physical target. In addition, the subjects were recorded with 2 different lighting conditions. The target's position, eyes position and screen coordinates are all available in 3D coordinates, which is used for the model created by Palmero et al.

Figure 6.9 illustrates the network, which predicts the direction a given individual is looking in [55]. The network takes a 3D space as input instead of the typical 2D images. Overall, it has a multi-modal recurrent convolutional neural network (RCNN) structure, the basics of which are described in Section 5.4.2. The structure is divided into four parts: normalization, individual, fusion and temporal.

First, the normalization part's goal is to align the angle of the head with the angle of the camera, as the gaze direction is not only affected by the eye position, but also the head position. This will reduce the appearance variability the network must learn to track. Finally, the normalization part crops the image into three sections: the face region, the eyes region and the face landmarks.

The second, third and forth part are all a part of the deep learning network. The system takes the three inputs streams through individual convolutional layers and fully connected layers, as shown in Figure 6.9. The individual streams are then concatenated and pass through two fully connected layers. The final, fourth layer is the recurrent part of the network. It uses information from previous time-steps to generate the final prediction after a last fully connected layer.

Overall, Palmero et al. proved the importance of including head positions in the model, as the model showed a 14.6% improvement on the EYEDIAP dataset compared to previous state of the art methods in 2018. The model showed most improvement for the images with large head movement, with a result of a 6.2 degree average error among participants. The result of the model given the images with minimum head movement is a 5.1 degree average error. However, these results are significantly worse than the results from the geometric model mentioned in Section 6.3.1. The geometric example had a mean error of 2.7 degrees, though this is mostly due to the difference of the restrictions. The models have a huge difference in the setup, whether the model is calibrated, and the general restrictions of the data types and allowed movements from the subjects.

6.4 Summary

In conclusion, the gaze tracking taxonomy introduced by Dario Cazzato et al. in 2020 has been explored through different examples within each of the categories *Scene Analysis* and *Beholder Eye/Face* [42]. By exploring each subcategory thoroughly, it seems apparent that research within Scene Analysis, and specifically third-person view, is very well suited for this project. Even though the gaze tracking techniques mainly strive to estimate the exact point of regard of individuals, this study will simply focus on estimating whether a point of interest is within the field of view of the individual people, so the margin of error is larger. Chapter 7 will look into an experiment using third-person view algorithms to solve the problem.

CHAPTER 7

Experiment

The development within gaze tracking technology allows for an estimation of what a person is fixated on. In this experiment the intention is to use gaze tracking to estimate what people are able to perceive from the outside world.

The experiments are based on the work by Dissing et al. described in Chapter 2 [8]. One of the work's shortcomings is the lack of a precise estimation of when a person is actually able to see the rearrangement of the cubes within the containers, as shown in Figure 2.3. The experiment in this study uses the same setup as Dissing et al. with the intention of improving the estimation of when a person is able to see the cubes being moved. Unlike other gaze tracking methods, this project focus on estimating a probability of whether the people within the images are able to perceive one specific target. Which objects within the image the people are looking at is thus not relevant unless it is within the boundaries of the target.

The chapter will initially describe specific technologies utilized for creating different methods to estimate a human's perception of a target within the scene. The technologies include a gaze tracking network, research within theory of mind, pose tracking networks, and face recognition techniques. The technologies are followed by a detailed description of 3 different methods for solving the problem. Finally, these methods are evaluated based on a set of videos designed specifically for this purpose.

The results and the experiment's shortcomings are discussed in Chapter 8.

7.1 Technologies

Various existing technologies have been utilised to create simple experiments for estimating people's perception of the world through gaze tracking. In addition to the gaze tracking model, the experiment also needs to be able to estimate the reliability of the result. This is done using probability distribution functions, which allows for a more precise estimation. The probability distributions are described within the specific method where it is used.

7.1.1 Gaze360

The setup has cameras viewing the scene from a third person view, which will be the main condition when selecting gaze tracking techniques. The camera setup makes scene analysis third-person view the most apparent building blocks for this study. Petr Kellnhoft et al. published their work with a third-person view called *Gaze360* in October 2019 [56]. The project focuses on estimating the 3D gaze direction of individuals in images, particularly when head is at a large angle resulting in only one or no eyes being visible from the camera's point of view. An example of these angles are seen in Figure 7.1.

In order to create the gaze estimation model, Petr Kellnhofer et al. had to create their own dataset to train the model. The new dataset is currently the largest publically available dataset of its kind when it comes to number of subjects and variety. The dataset contains both indoor and outdoor photos, as well as photos where the head is at a strong angle to make it possible for the model to handle all natural scenarios. The data is acquired by using a Ladybug5 360 degree panoramic camera. Multiple subjects within the camera view look at an AprilTag, which in turn makes it possible to translate their gaze

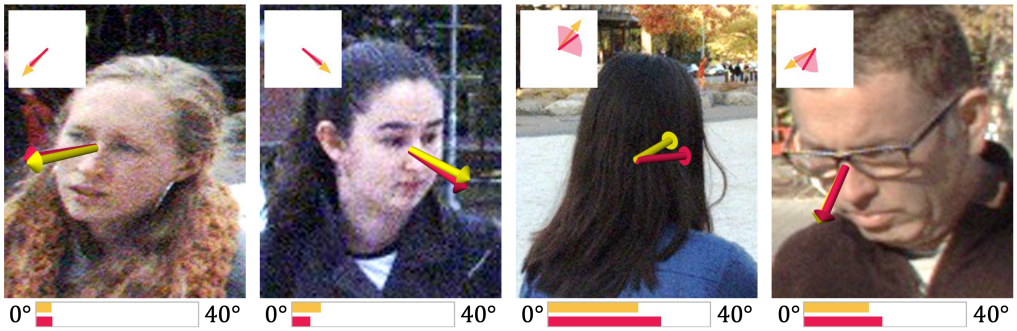


Figure 7.1: Results from [56]. The yellow error is the ground truth while the red is the Gaze360 prediction. The bars underneath show the actual error (yellow) in degrees and the estimated error (red).

direction towards to the target [56]. In addition to static images, the dataset also contains annotated video footage.

The Gaze360 model is designed to predict the 3D gaze direction resulting in two outputs. The outputs are the predicted gaze direction, (θ, ϕ) and an appropriate offset able to calculate the confidence level as an error quantile estimation, (σ). Overall, the network is built up on a video-based gaze-tracking model using bi-directional Long Short-term memory capsules (LSTM) [56]. The bi-directional LSTM creates a dependence on both previous and future inputs to utilise a gaze that is a continuous signal. The model is trained by having 7 frames as input. However, the design can handle a single frame as well, if this might be the case.

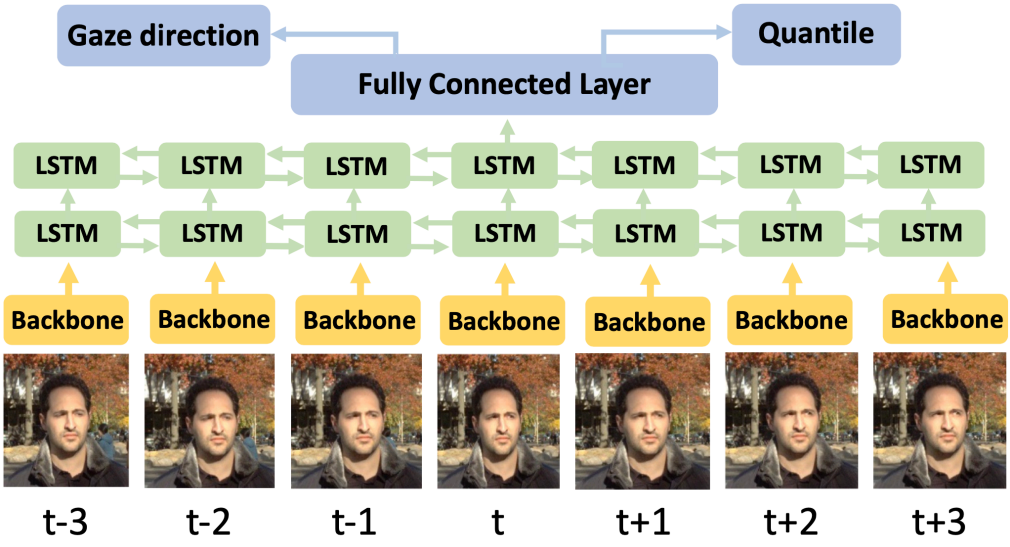


Figure 7.2: The architecture of Gaze360 model from [56].

The architecture of the model is illustrated in Figure 7.2. The image input is simply the cropped head which is first processed by the backbone. The backbone is an ImageNet pretrained ResNet-18 to extract the high-level features of the faces. The backbone is followed by the bi-directional LSTM network, and finally a fully connected layer to produce the outputs. The gaze direction output is spherical coordinates as it is believed that these are more naturally interpretable [56]. The error quantile estimation is calculated using a pinball loss function. The error quantile is the estimated offset from the predicted gaze direction to reach the 10% quantile and the 90% quantile, as seen in Equation

7.1.

$$q_{10} = (\theta - \sigma, \phi - \sigma) \quad \text{The 10\% quantiles} \qquad (7.1)$$
$$q_{90} = (\theta + \sigma, \phi + \sigma) \quad \text{The 90\% quantiles}$$

The evaluation of the Gaze360, in comparison with other networks, show a state-of-the-art result of a mean angular error of 13.5 degrees within a field of all 360 degrees. This proves its efficiency when allowing people to look in all angles. The network was also tested on a vending machine setup, where the results showed an accuracy of 68%. This accuracy is quite small in comparison to other gaze tracking technologies, which can have an angular error of only 1%. However, given the fact that Gaze360 has no calibration and allows for the participant to move head, body and eyes freely within the image, it becomes a unique solution.

7.1.2 Theory of Mind

When teaching a robot theory of mind, as the implementation by Dissing et al. [8], it is essential for the robot to be able to detect when a person gains a false belief. The critical idea behind this work is that the robot will be more aware of the circumstances if it has a theory of mind, which in turn will create a better social acceptance.

But to be able to be aware of circumstance-appropriateness, it is also essential to understand what people actually expect the robot to believe and understand. An example would be whether a person assumes a robot would recognize a change of event if a blanket was thrown over its face. To gain circumstance-appropriateness and familiarity means that the robot will know and react as people expect.

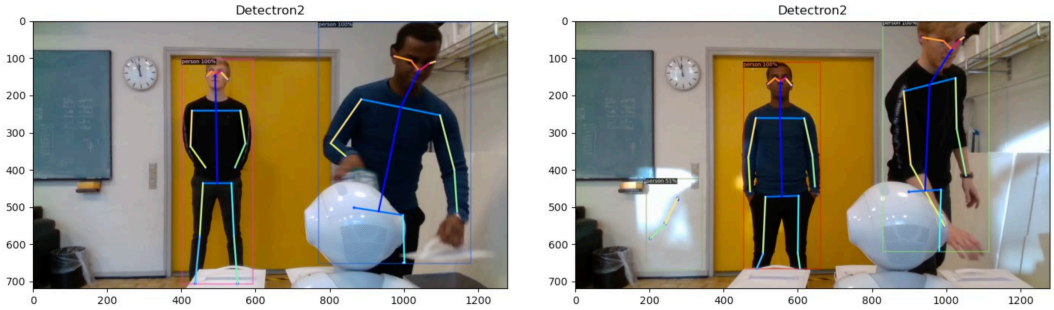
So, how do people actually expect a robot to act? Thellman and Ziemke created a number of tests to analyse what people expect a robot to know based on false-belief tasks [57]. They used the same robot as Dissing et al., a SoftBank Robotics Pepper. The subjects where asked to predict whether the robot would gain a false belief when a transparent or opaque curtain was drawn in front of it. The results showed that people do infer perceptual beliefs onto a robot with a tendency to imply anthropocentric perceptual capabilities. Evidently this also implies that it is more difficult for people to track a robot's beliefs and actions when they differ significantly from humans' abilities. Thellman and Ziemke also found that people can learn to predict a robot's actions after familiarising themselves with it during prolonged exposure, while simple verbal instructions weren't particularly efficient.

So, for people to easily understand the beliefs of a social robot and predict its actions, it becomes apparent that its capabilities need to be similar to humans. When people naturally assume robots act with the same capabilities as humans, they will become more accepting if the implementation reflects this.

7.1.3 Extracting Features

The Gaze360 network needs to receive 7 ordered cropped image of the face of the person of interest. In order to detect people and their features within a video, the deep learning network Detectron2 by Facebook has been utilised [58]. Detectron2 is the second version and based on the original project Detectron. It contains a *Model Zoo*, which contains a set of different pre-trained checkpoints. One of these Model zoo checkouts is *keypoint RCNN R 50 FPN 3x*, which is able to determine the location of 17 keypoints on the human body. These include ears, eyes, nose, shoulders, elbows, wrists, hips, knees and feet.

Figure 7.3 shows an example of the detrectron2 keypoints being used to locate two people in an image. In addition to the keypoints, it also outputs a probability scale of how sure the network estimates a person has been located. The two people in Figure 7.3a are both 100%, but in Figure 7.3b it estimates the possibility of the shadow being a person with 65% accuracy.



(a) Detectron2 locating people in a frame with a cofidence of 100%.

(b) Detectron2 locates two people and a shadow in the frame. The shadow has a confidence of 65%, but is not a person.

Figure 7.3: Screenshots of implementation of detectron2 [58] on a video taken from Test 3.

7.1.4 Face Recognition

Another main component is keeping track of the different people. Gaze360 has an input of 7 frames: 3 frames before, 1 frame it wants to analyse and 3 frames after. This makes it essential to keep track of the people throughout the frames, as feeding an image with a different person would not result in an accurate result.

To recognize the people detected from detectron2, a face recognition library by Adam Geitgey is utilised [59]. The model has an accuracy of 99.38% on the Labeled Faces in the Wild benchmark [60], which consist of 13233 images of 5749 people.

However, the face recognition is not always able to locate all features of the face, if the face is at an angle, which detectron2 will localize. For this reason, a hierarchy of techniques has been implemented to ensure that people are recognized correctly. The other technique used is based on the fact that the data is a video. Therefore, a person will not be able to move very far from one frame to the next. The location of the person in the image is therefore used to determine who the person is given the following emphasises.

1. **Location and recognition** both have to result in the same identity for the person of interest.
2. **Location** is used when the face recognition and location do not give the same result.
3. **Face recognition** by itself is the last attempt to identify the person if the other solutions failed.

The hierarchy structure allows for identifying a person given all criteria and utilises the knowledge common within video frames.

7.2 Methods

Three types of methods are implemented to determine whether the people visible in videos are able to see a specific target object. The methods are all based on Gaze360, as it consistently includes a gaze direction in every frame. In addition, Gaze360 also presents a confidence measurement, which allows an analysis of the result in a probable manner.

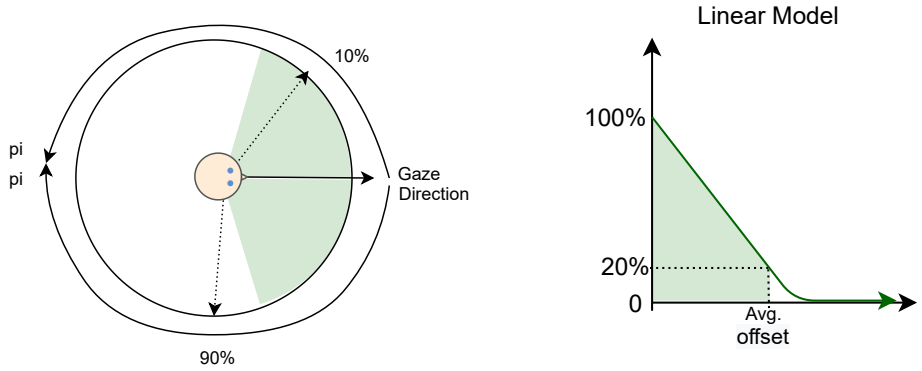
For scoping the experiments, the gaze vectors have been converted to 2 dimensions instead of the direct output in 3 dimensions. In addition, the target location will not change throughout the frames, but the implementation is created in such a fashion that the bounding box of the target location can be passed into the code, and thus change based on the object.

7.2.1 **Method 1**: Facing the Camera

The method used for estimating what an individual is able to see in the implementation of theory of mind by Dissing et al. is simply a detection whether the person is looking at the camera or not. If the person is facing the camera, it is assumed that this person is able to see the target with 100% accuracy. In contrast, when the person leaves or turns around, the method is 100% sure that the person cannot see the target. This method simply uses detectron2 keypoints to detect people, and thus recognises whether they are facing the camera using face recognition. The facial landmarks can only be located when the whole face is visible. The implementation is therefore independent from the target's location and the gaze direction.

7.2.2 **Method 2**: Mean gaze direction within target bounding box

The second method uses gaze360 and the bounding box to estimate the visibility of a target to the subject. The intention is to create a linear model that estimates the probability dependent on the output from Gaze360. The linear model is calculated based on the gaze direction having a value of 100% at 0 degrees, and the average of the quantiles angles from the gaze direction have the value 20%. This is illustrated in Figure 7.4, where Figure 7.4a illustrate the quantiles and gaze direction output from Gaze360 in relation to the angles, and Figure 7.4b show the linear interpretation of the model.



(a) Illustration of a linear model based on Gaze360 outputs. The actual orientation of the angles are dependent on the camera's point of view.

(b) A linear model calculated based on the average of the angluar offset from the quantiles to the gaze direction.

Figure 7.4: Illustrations of how the linear model is created using an interpretation of gaze direction values as parameters.

The offset angles are not linear. However, the simplification creates an easy and fast mapping which utilises both the offset and the gaze direction. The model is then used to create a heatmap onto the image, so every pixel in the image gets an estimated value on the basis of the linear model. An example of this heatmap is shown in Figure 7.5, where the red line is the estimated gaze direction, the orange and pink lines are the quantiles, the green lines are the angles towards the target, the white mesh is the estimation, and lastly the blue bounding box is the target. The angles are dependent on the camera angle and is thereby neither completely horizontal nor vertical.

The estimation of how probable the person in the image is to look at the target is simply based on the mean of all the pixels within the target bounding box. So, the example in Figure 7.5 is 0.00% because the majority of the heatmap values within the blue box is 0, while a few might be around 0.1 as the heatmap touched the edge.

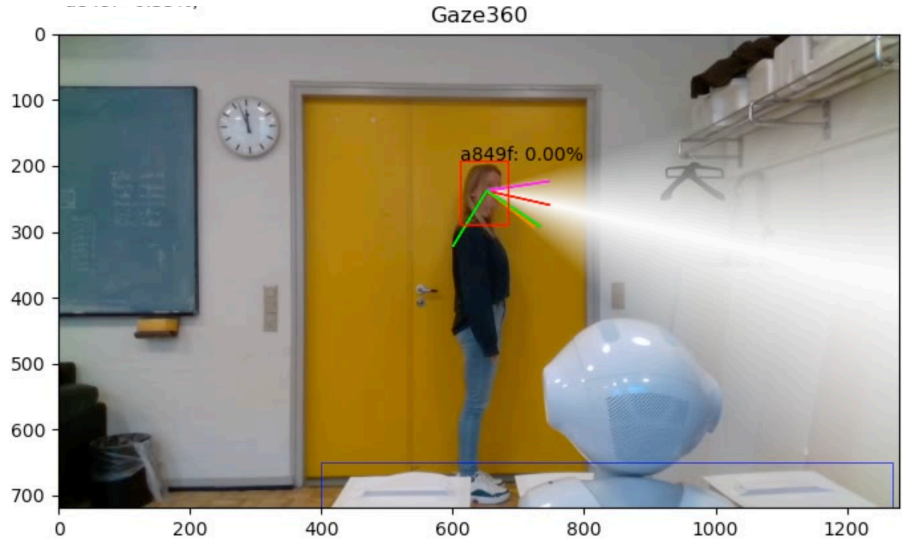


Figure 7.5: Example of the heatmap created from the linear model. The image is from the analysis in Test 1, which is introduced in Section 7.3.

7.2.3 **Method 3**: Von Mises distribution based on angles

The third and final method is designed to utilize the distribution knowledge from Gaze360 by imitating a normal distribution in circular statistics.

The offset given by Gaze360 is deemed to be the 10% quantiles and 90% quantiles as illustrated in Equation 7.1. The offset quantiles simulate an isotropic distribution, which is a simplification scoped by the authors. An isotropic distribution is a multidimensional Gaussian distribution with the covariance matrix being a scalar variance multiplied by an identity matrix.

The offset from 3D to 2D is no longer symmetrical, as the gaze direction typically won't align with the 2D plane. To utilize the distributional knowledge, the initial intention was to calculate the distribution as a *skewed normal distribution*. The skewed normal distribution is not symmetrical, which the two quantiles in 2 dimensions are not either. However, the skewed normal distribution proved complicated, as three parameters needed to be calculated numerically and the axis was based on angles and not linear data.

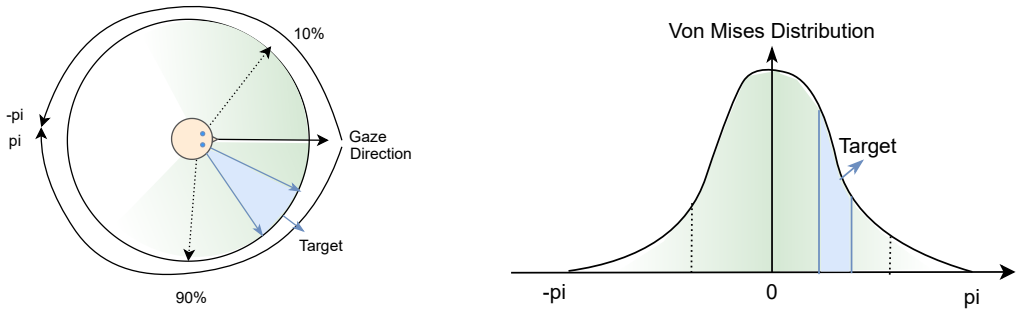
Setting the mean gaze vector to the angle 0 and using the differences from the mean and to the quantiles' angles, three angular data points are created. These data points describe the mean, the 10% quantile and the 90% quantile within a circular statistical distribution. Within circular statistics, the analogous to the normal distributions on the line is the *Von Mises* distribution [61]. The solution therefore uses a Von Mises probability distribution to simulate a gaze distribution function. The formulae for the Von Mises distribution is seen in Equation 7.2.

$$f(x \mid \mu, \kappa) = \frac{\exp \kappa \cos x - \mu}{2\pi I_0(\kappa)}, \text{ where } \frac{1}{\kappa} \propto \sigma^2 \tag{7.2}$$

The only unknown variable in the distribution is kappa (κ), which is calculated numerically using *scipy.optimize* functions. The Von Mises distribution is symmetric, so the skewed quantiles cannot both influence the algorithm. The distribution is thereby built up based on the largest uncertainty, and thereby the numerically largest of the two quantiles' angles (named a in Equation 7.3). Kappa can thus be calculated based on the fact that the mean (μ) is 0 and cumulative Von Mises distribution function is equal to 0.9 when integrating from 0 to a, as shown in Equation 7.3.

$$0.9 = \int_0^a f(x \mid 0, \kappa)dx, \text{ where } \kappa > 0 \text{ and } a = max(\mid a_{10} \mid, \mid a_{90} \mid) \tag{7.3}$$

Equation 7.2 is illustrated within the gaze tracking scenario in Figure 7.6. The intention is to use the distribution to calculate the integration within the angles from the mean to the target's corners, which allows for a probability-type estimation.



(a) Illustration of a Von Mises model given the Gaze360 output.

(b) The Von Mises model calculated based on the maximum of the angular offsets given by the quantiles to the gaze direction.

Figure 7.6: Illustrations of how a Von Mises model is used based on gaze direction values as parameters.

7.2.3.1 **Method 3a**: Based on extended angles

A final experiment was implemented with a minor change to the method using Von Mises distribution. The small change is related to the angel towards the target, which the integration is based on. The target angles are in this scenario extended by 30 degrees, so the integration will have a minimum of 60 degrees, if the target's angles were 0. The extension is based on the assumption that people are able to see movements within the peripheral field of vision. If a person's gaze is just outside the target, with a great confidence, the target might still be within the peripheral field, or even paracentral view, but the simple Von Mises model will not take this into account.

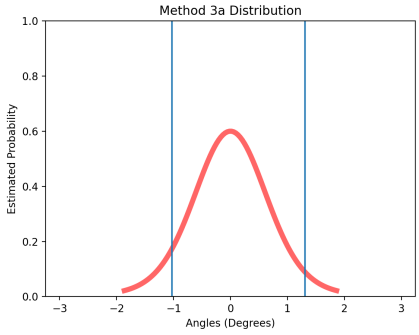


Figure 7.7: An example of the Von Mises distribution for method 3a. The vertical blue lines are the angles towards the target.

An example of the Von Mises distrubtion for method 3a shown in Figure 7.7. The vertical lines are the angles from the gaze direction, from Gaze360, and to the target's outer edges. In this example it is clear that the participant is able to see the target.

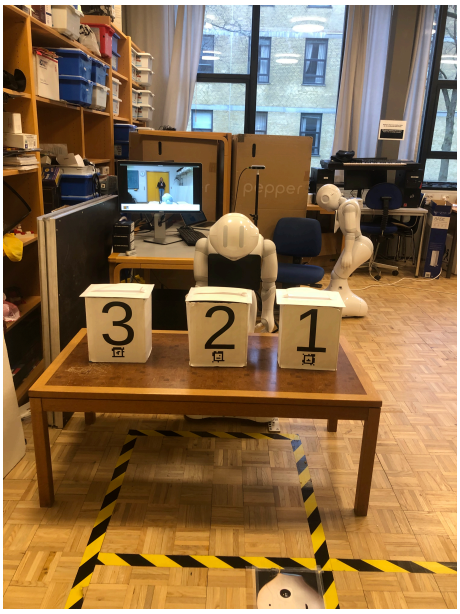
7.3 Implementation and Tests

The implementation of the methods are written using python 3 on a MacOS operational system. The neural networks are able to run using the graphic processing unit. However, the cuda implementation is not able to access the internal GPU of the Macbook Pro. All calculations are thereby run on the CPU available. The solutions have been implemented using the PyCharm IDE. The implementation uses several publically available python libraries to create the graphics etc.

7.3.1 Test Data

To test the solutions, test data was gathered based on 3 different test scenarios. The test scenarios all have the same initial setup, using the same tools as in the paper by Dissing et al. [8]. The setup is shown in Figure 7.8. The test scenes are filmed using intel Realsense depth camera D435i, but the depth in the output has not been used in order to scope the project to 2D.

(a) The other side of the scene, where the camera is placed to record. The camera is exactly 1m from the yellow and orange tape, and thus 1.7m from the closest distance.



(b) The view the participants will have when completing the tasks. Each square of tape is 70cm, and is thereby used to determine how far a participant is from the camera.

Figure 7.8: The setup for the tests estimating whether people are able to perceive actions on the table with the containers. The setup is originally created by Dissing et al. [8].

For the first test plan, the subject must always look at the target, and feel able to observe any changes occurring. The subjects are positioned so they face to the left, front or right, at three different distances from the camera. This allows for at least 9 different types of video clips, where the result is assumed for the subject to always see the target. The target is the combined three containers on the table.

The second test is similar, but the subject must never look at the target. So the expected results from test 2 is for the subject never to be able to see the containers. The third and final test is a story line, to understand how the solution works within a false-belief task. The subjects are therefore asked to place a cube in one container, then have one subject look away, and thereby not see the other subject

moving the cube to a different container. The test plans have all been described in greater detail in Appendix B.

The first two tests are completed by the same subject, but with or without glasses. Test 1 at the distance of 310cm have also been completed by two young male participants, who also conducted the third test. Overall, the test resulted in 23 videos of approximately 15 seconds each.

7.3.1.1 Pre-possessing

To analyse the different videos thoroughly, the videos have been edited. The initial and final $0-3$ seconds have been removed, as these were used to get into or out of position. In addition, each video has been divided so it only contains one orientation instead of all $3-4$ depending on the test.

The results presented in Section 7.4 only use the videos from test 1 and test 2 where the orientations are either *left*, *front* or *right*. The clip where the individuals are facing away from the camera have been removed, as test 1 is not possible from this orientation. The distributions of the test 2 facing away from the camera is included in Appendix C.

7.4 Result

The data gathered from the test is used to analyse 3 different methods of estimating individuals' ability to perceive an object. For every second frame in a test video the script is run once. This section will present the distributions of the saved results for each method. The results shown below are only an extraction from all the results, which are presented in Appendix C.

The term *probability* is used frequently to describe the estimation of how confident the algorithm is that a person can see the target. As the data is based on a neural network, which in turn has had a dimensional reduction, the more correct phrase is *estimated probability*, as the estimation might not reflect the terminology accurately.

7.4.1 Method 1 Results

The results from method 1 are estimations of how likely people are able to see the target in every frame. So, for every video in the test data, each detected person within a frame will have an estimated probability. The general statistics based on these estimations are written in Table 7.1. The median, min and max values in Table 7.1 are all either 0.0 or 1.0, which is a result of method 1 being binary.

Test	**#Estimations**	**Mean**	**Median**	**Min**	**Max**
Test 1	698	0.6	1.0	0.0	1.0
Test 2	774	0.26	0.0	0.0	1.0

Table 7.1: Statistical results from Method 1. The table shows which test data the values are based on and the number of estimations within each test dataset. In addition, the mean, median, min and max value of the estimations are shown.

Test 1 is based on the data where the participant looks towards the target in every frame, while test 2 is when the participant never looks at the target. The overall distribution of the results from test 1 and test 2 using method 1 are shown in Figure 7.9.

Figure 7.9 shows that both distributions are quite divided, which is caused by the binary data. The distribution therefore resembles a Bernoulli Distribution, as it simply has successful or failure cases. Test 1 has 424 estimations equal to 1.0, while test 2 only has 200. To understand whether the distributions are statistically different, a Mann-Whitney u-test has been run giving a statistics u-value

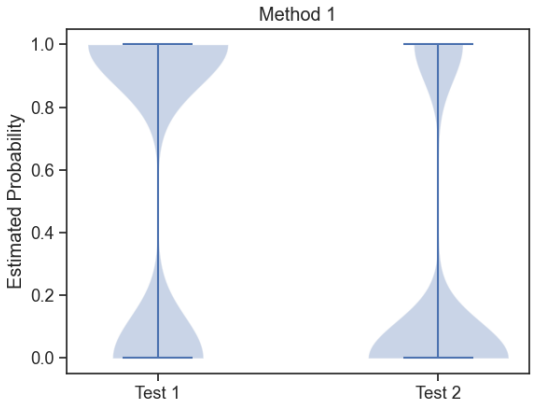


Figure 7.9: The results based on method 1 on test 1 and test 2 (with the exception of the participant facing away from the camera in test 2).

of $1.76 \cdot 10^5$ and a p-value of $5.42 \cdot 10^{-42}$. The u-value is the minimum of two u-values calculated for each distribution, where the u-value is found based on a point-system in a Mann-Whitney u-test [62]. The p-value is below 0.05, which means it is robust to assume that the results from test 1 and test 2 are statistically different.

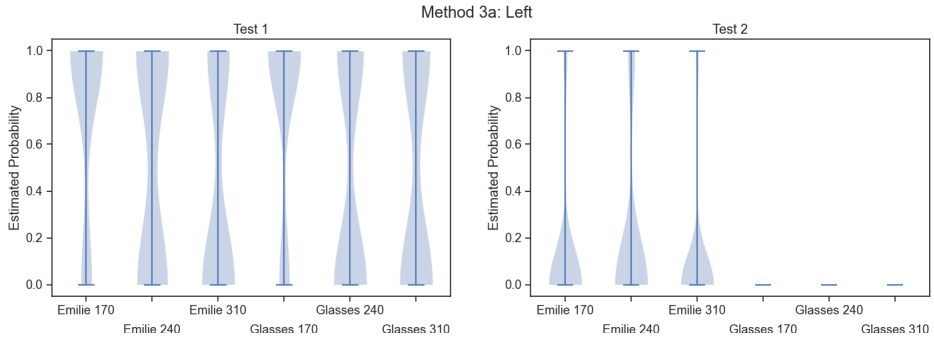


Figure 7.10: The results from method 1, given the left orientation and different distances and people. The cases named *Emilie* are the same test subject as *Glasses*, but without glasses. The numbers define the distances, so e.g. *_170* is at a distance of 170cm from the camera.

The test data was further divided into different orientations and different distances. The results of one of these variations is shown in Figure 7.10, which shows the subjects facing left. For test 1 the results are quite evenly spread between 0 and 1 with an exception of the two cases *Emilie_170* and *Glasses_170*. The two exceptions are both from a distance of 170cm from the camera, which seems to estimate that they look at the target more frequently. In comparison, the results for the left orientations within test 2 are also quite interesting. None of the cases where the subject wears glasses are above 0, and the remaining cases only have few elements above 0. This shows how accessories, such as glasses, can cover the face to such a degree that the face landmark cannot be located by the face recognition algorithm.

7.4.2 Method 2: Mean gaze direction within target bounding box

Like method 1, the general statistical values are calculated given the estimations from method 2. These values are described in Table 7.2. Unlike method 1, the results are no longer binary as seen in the

maximum values of 0.96 and 0.97. Based on the statistical values presented in Table 7.2 only the mean vary with 0.16 for test 1 and 0.06 for test 2 in addition to 0.01 difference for the maximum value.

Test	#Estimations	Mean	Median	Min	Max
Test 1	698	0.16	0.0	0.0	0.96
Test 2	774	0.06	0.0	0.0	0.97

Table 7.2: The mean, median, min and max are calculated based on estimations of how probable an individual can see a target given my method 2. #Estimations are the number of estimations given the test sets.

The similarity is also reflected in the overall distribution, as both cases have the main portion closer to 0.0 with a slightly larger tail upwards for test 1. A Mann-Whitney u-test is used to determine whether the estimations from test 1 are actually statistically different from test 2. Based on the null hypothesis that the estimations are from the same distribution, the statistics being a u-value of $2.19 \cdot 10^5$ and a p-value $5.84 \cdot 10^{-16}$. As the p-value is quite a bit less than 0.05, the null hypothesis is rejected and the estimations from test 1 and test 2 are not from the same distribution. Thus, even though the distribution in Figure 7.11 seem quite similar, they are not. Based on a threshold of 0.5 test 1 computes 230 estimations above the threshold and test 2 result in 131 estimated probabilities above 0.5.

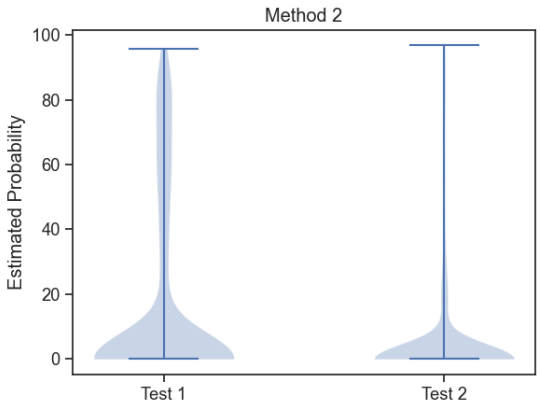


Figure 7.11: The results based on method 2 on test 1 and test 2 (with the exception of the participant facing away from the camera in test 2).

An example of an estimation of a frame is included in Figure 7.12. The example demonstrates how the probability is calculated by taking the average of the heatmap values within the target bounding box. The values from the heatmap are based on the offset angles from the Gaze360 estimation. Thus, the algorithm estimates how much of the bounding box the heatmap covers.

The distributions for the individual orientations are shown in Appendix C. Overall, all the trends seem similar to Figure 7.11 with the results in test 1 being slightly higher than test 2.

7.4.3 Method 3: Von Mises distribution based on angles

The general statistical results for using method 3 are shown in Table 7.3, which, unlike the previous, results have a larger visible difference between test 1 and test 2. Both the mean and the median reflects how test 1 are given higher estimations than for test 2. As the median is smaller than the mean in test 1, the majority of the data is closer to 0 and the distribution is not symmetrical.

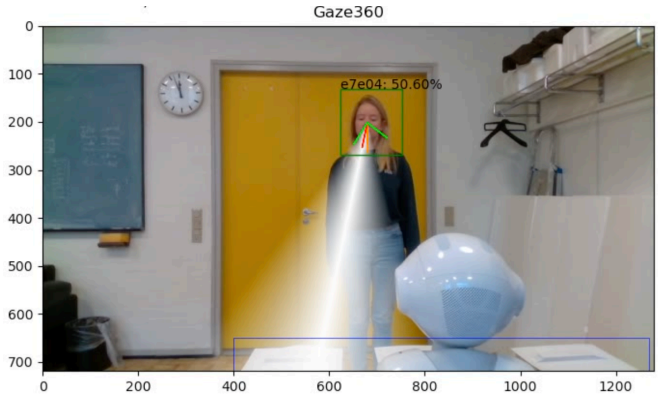


Figure 7.12: The frame is from the test scenario where the subject does not wear glasses, at a distance of 240cm from the camera and facing forward. The algorithm run is method 2.

Test	#Estimations	Mean	Median	Min	Max
Test 1	698	0.46	0.35	0.0	1.0
Test 2	774	0.06	$3.10 \cdot 10^{-4}$	0.0	0.88

Table 7.3: Statistical results for the test data using method 3. The mean, median, min and max describe the estimations for the given data sets, while #Estimations describe the number of estimated data point within each set.

Like the previous models, a Mann-Whitney u-test is conducted giving a statistics u-value of $1.08 \cdot 10^5$ and a p-value of $1.5 \cdot 10^{-88}$, proving a statistical difference between the estimations for test 1 and test 2. This discovery is also seen in Figure 7.13, which illustrates the two distributions. A large amount of the estimations for test 1 are close to 1.0 in comparison to test 2, where the estimations never reach a result of 1.0 as the maximum value within the estimation is only 0.88. Creating a threshold of 0.5 the estimations for test 1 result in 309 estimations, while test only result in 9 estimations above 0.5.

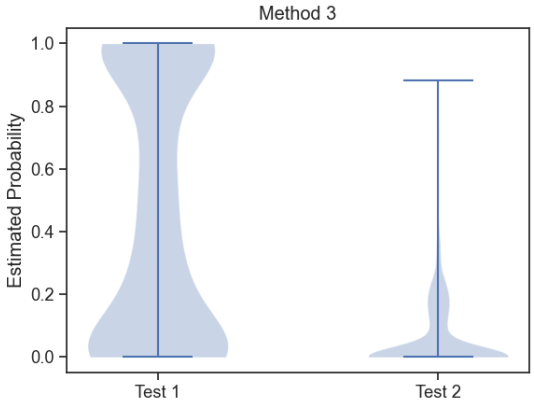


Figure 7.13: The results based on method 3 on test 1 and test 2 (with the exception of the participant facing away from the camera in test 2).

Figure 7.14 focuses on the individual cases when the subject is either directly facing the camera or facing to the right. When the individual is facing forward, the results for test 2 are mostly below 0.2

while test 1 almost have no values less than 0.2. There is a moderate trend for the cases in test 1 with glasses to be closer to 0 than the cases without.

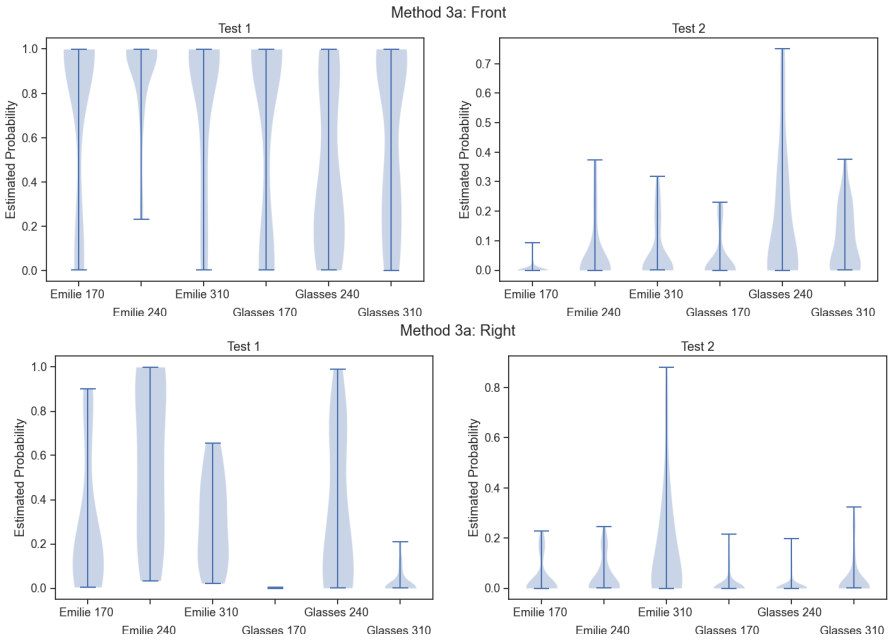


Figure 7.14: Distributions from analysis with method 3 when the subject is facing the camera or facing right.

In addition, for the right orientation, the distribution for test 2 seems similar to the front facing. However, the same is not the case for test 1. Test 1 for the right orientation is not as confidently close to the value 1.0, but actually closer to 0.0. One case where the subject is close to the camera and wearing glasses, *Glasses_170*, no estimations were found above 0.0.

7.4.3.1 Method 3a: Based on extended angles

The final method implemented is based on the same calculation as method 3, but with the target's outer angles extended by 30 degrees. The results reflect this extension, as all the statistical values for the datasets have enlarged. The maximum value for test 1 is now also 1.0, while the minimum value is very close to 0.0, but never reaches the exact value. The median is the largest change from method 3 to method 3a, with a percent deviation of 174% for the difference of median for test 1 and a percent deviation of 499% for test 2.

Test	**#Estimations**	**Mean**	**Median**	**Min**	**Max**
Test 1	698	0.75	0.96	$4.09 \cdot 10^{-13}$	1.0
Test 2	774	0.23	0.15	$-1.43 \cdot 10^{-13}$	1.0

Table 7.4: Statistical results for the test data using method 3a. The table show close to a 0.0 minimum estimation within both test datasets, a 1.0 maximum value, and varying means and medians. Test 2 has a few more data points than test 1.

The distribution of the estimated values overall is shown in Figure 7.15, where the majority of the values are either close to 1.0 and 0.0 for test 1 and test 2 respectively. The Mann-Whitney u-test proves the distributions to be statistically different with the statistics u-value being $6.42 \cdot 10^4$ and the p-value

$1.70 \cdot 10^{-141}$, which is less than 0.05. Based on a threshold of 0.5, test 1 results in 539 values above the threshold and test 2 results in 105 values.

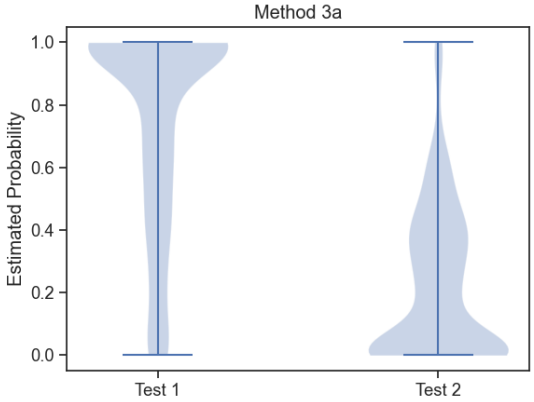


Figure 7.15: The results based on method 3a on test 1 and test 2 (with the exception of the participant facing away from the camera in test 2).

The right orientation given method 3a is shown in Figure 7.16. Test 1 data is distributed with a trend towards 1.0, with an exception of *Glasses_170*, while test 2 has a trend towards 0.0. The case with *Glasses_170* does have a large portion within 0.0, but is quite evenly distributed. The results from test 2 have moved more towards the middle area in comparison to test 2 for method 3, where it was closer to the bottom.

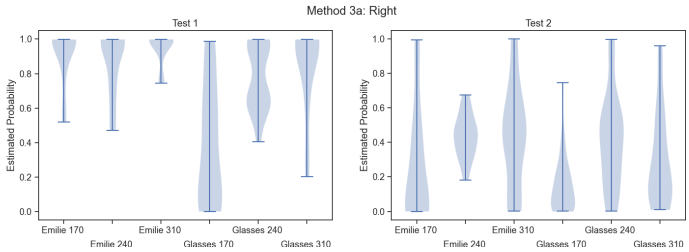


Figure 7.16: The detailed analysis when the subject is facing right. The estimations are based on the implementation of method 3a.

7.4.4 Test 3

The third test dataset created is based on a story telling environment. The videos contain two people instead of only a single person, and thus test the solution's ability to distinguish between the people and estimate their ability to see the target individually. Figure 7.17 includes three frames from a video where the estimation of whether the people can see the target is based of method 3a.

A set of 8 videos are estimated in total for test 3 for method 3 and 3a, where the participants were asked to look away facing different orientations, and switch with one another. These videos can be found using a link in Appendix D. Test 3 ensures that the identification algorithm could recognize multiple people. The results shown through all the videos prove that the algorithm, in these cases, are able to recognize two people and give them individual estimations.

Initial results did show identified people being switched or given new identities. These error videos are also included in Appendix D.

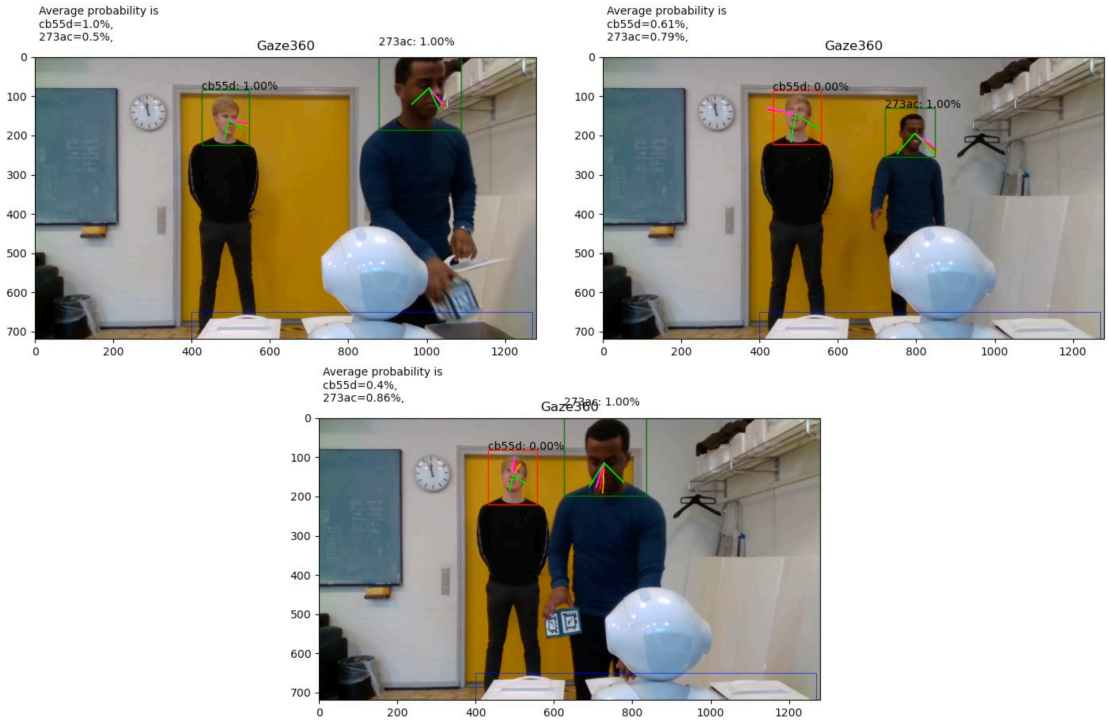


Figure 7.17: Results from the test using method 3a, where one person is instructed to be facing forward, but look away the second time the other person moves the cube. The average probability in the upper left corners of the frames is the average of how much the individual looks at the target based on the previous frames. The probability value above the participants' heads is the estimated value in the frame and the numbers are the persons' random ID. The lines indicate where Gaze360 estimates that the subjects are looking. Lastly, the bounding box around the participants' faces are green when the estimation is above 0.25 and red when it is below.

7.4.5 Computational Cost

Given the lack of access to a compatible graphics processing unit, the computational cost of running the methods was expected to be quite large. On average, the computational cost of analysing a video of approximately 4 seconds takes an average of 5.3 minutes. Of this cost, only 0.3 minutes were used by Gaze360, and for method 3 and 3a only 0.22 seconds on computing the distribution. This shows how the true cost inefficiency might not only be due to the lack of a GPU, but that other parts of the implementations might be at fault. The computational cost is not measured for extracting features using detectron2, face recognition algorithm, processing of the video and the generation of the output video and json file. These all contribute to the cost, but the amounts are unknown at this stage. The precise data for the cost given the different methods is shown in Appendix C.

7.5 Summary

An experiment testing 3 different methods for estimating people's perception of a specific target have been implemented and evaluated. The evaluation is based on videos from a setup similar to the work by Dissing et al. [8]. The overall result showed a greater degree of separation between the videos where the subject was looking at the target or not. The distributions of the results for all the different parameters are described in Appendix C. All methods have a Mann-Whitney u-test p-value under 0.05 indicating a statistical difference between the two tests. The videos showing the estimations for all test datasets,

including test 3, are linked in Appendix D.

The results from the experiments are discussed in Chapter 8, and potential improvements are described in Chapter 9.

CHAPTER 8

Discussion

Throughout this report various works try to understand how humans perceive the world has been introduced. This shows how perception cannot simply be estimated within one field of study, but needs a mixture of various works within cognitive psychology, neuroscience, computer vision, deep learning etc. The experiment conducted in Chapter 7 emphasises peoples' use of eye movement to fixate on attentive stimuli. This chapter discusses how all these fields relate to the experiment and introducees shortcomings to the implementation based on gaze movements. The results from Chapter 7 are initially discussed and compared to each other. Then the solution is evaluated in perspective of the research within cognitive psychology and the additional fields.

8.1 Results from Experiment

Some of the results from Chapter 7 are shown again in Figure 8.1, which allows for a more simple comparison. This section will first interpret and discuss the results independently for method 1, and then continuously discuss the methods and compare them to the previously discussed methods.

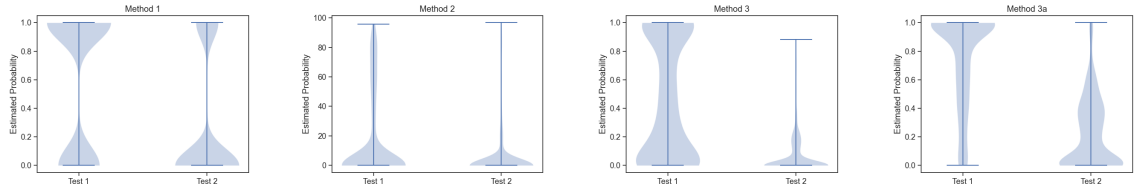


Figure 8.1: The Figures show 4 implemented methods that estimate how likely a user is to be able to see a target. Method 3a uses the same implementation as Method 3, but with the outer angle of the target being extended.

8.1.1 Method 1

The first model is based on neither gaze direction nor target location. This is reflected in the results, as the estimation does not look like the ideal solution. The distributions for test 1 and test 2 are statistically different, but the optimal result is having all estimations in test 1 be close to 1.0 and all estimations from test 2 close to 0.0. In some cases this method does seem quite capable i.e. when the participants are facing left or right. In these scenarios the results for test 2 are very close to 0.0 as seen in Figure 7.10 or Appendix C.

Based on the research, this technique might not be as overly simplified as first thought. The preferred viewing area for humans is within the 30 degrees in front of our face, as described in Section 4.2.1. In the scenario where the individual is facing away from the target, the nasal portion limits the eye movement to approximately 60 degrees. E.g. if a person facing 60 degrees to the left and looks straight ahead, the area of the target will only be visible within the far peripheral view. So, by only including the images where the user face the camera, one ensures that the target should be within the person's preferred viewing area. However, eye movement does allow an individual to move their gaze fixation without

head-movement and thus has the target within the para-central field of vision. When the face is at a large enough angle so that the algorithm cannot detect the facial features, method 1 will estimate that the user cannot see the target. However, as eye movement is possible, the target can be within the user's field of vision, but method 1 does not take this into account.

Section 4.2.1 describes how people prefer to move their head when looking in a direction larger than 35 degrees to either side, but it is still possible to gain a perception of a target within the peripheral field. Studies have shown that people are good at interpreting movement in the peripheral field. However, the far peripheral field does have fewer photoreceptors, so the likelihood of perceiving any details is lower.

In addition, method 1 does not take the location of the target into consideration. In this test, the target was located in the bottom right corner and thus in the foreground of the image. However, if the target was not placed in front of the camera, the solution would no longer be viable, as the method has the assumption that the target of interest is placed close to the camera.

Given the assumption that the target is very close to the camera, and that people will not intentionally trick the robot, by only looking at the target with the far corner of their eyes, this simple methods does seem able to estimate whether a person is able to see the target. However, the test datasets do include these gaze scenarios where the target is not visible within the preferred viewing area. This is reflected in the results as method 1 was able to estimate 424 correct in Test 1, which is 60.7% correct. The results for test 2 show a 74.1% correct estimation at 0.0. Overall, the combined percentage correct for both test 1 and test 2 is seen to be 67.8% calculated in Equation 8.1

$$\frac{424 + (774 - 200)}{774 + 698} \cdot 100 = 67.8\% \tag{8.1}$$

8.1.2 Method 2

To understand the results from method 2, two examples from the estimated probability have been included in Figure 8.2. The two examples demonstrate how the average heatmap within the target's bounding box vary based on the offset angles from the Gaze360 estimation. When the confidence is small, which results in a large offset, the heatmap becomes more spread. In comparison, when the confidence is high, the heatmap is much narrower, and therefore do not reach as much of the bounding box.

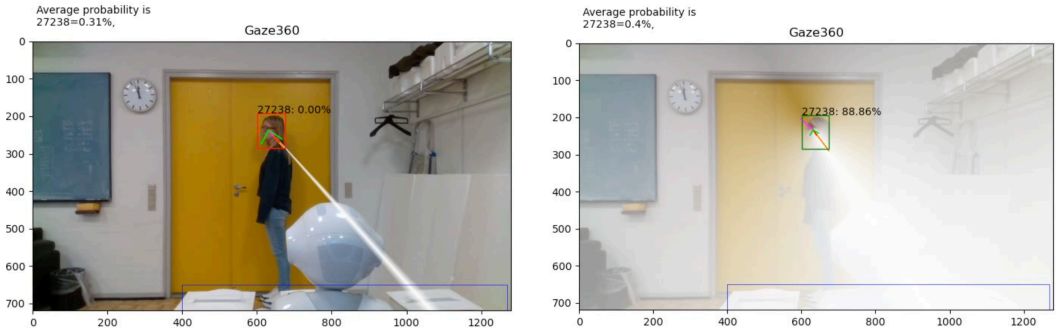


Figure 8.2: Two frames from the test scenario where the subject wears glasses, at a distance of 310cm from the camera and facing left. The algorithm run is method 2.

This is the exact opposite of what the intention is for this solution. If there is a high confidence in the individual looking at the target object, the resulting estimation should also be higher. This error is mainly due to the calculation of the heatmap being linear with an intersection always set to 100 at the 0 angle. The intersection value should decline when the offset is larger to ensure a more reliable

distribution. The linear model may simply be too simplistic a method. The photoreceptors inside the eyes are not distributed like a linear function, but have the density becoming slowly more sparse.

Given the results in Figure 8.1 method 2 is not able to distinguish between the datasets quite as well. A slight difference might be detectable when the participant is facing the camera. However, the remaining cases do not differentiate much. Given a threshold of 0.5, only 33.0% cases were correct for test 1 and 83.1% for test 2. It seems that most of the estimated probabilities are simply 0.0 making test 2 score higher. Finally, this gives a total correct percentage of 59.3%.

A small experiment, which changes the threshold from 0.5 to 0.25, has shown to make no big difference. This simply resulted in 1 more correct value for test 1 and 4 less correct values for test 2. Method 2 is therefore quite distinctly over-simplified as it only estimates a large probability when the angular offset is large.

Compared to method 1, method 2 seems less capable of predicting people's perception of the world. If the linear model was exchanged to a more complex model, the results may have been different.

8.1.3 Method 3

The main concept errors in method 2 are remedied in method 3, as the distribution of estimated probabilities based on Gaze360 are no longer linear. Having the distribution as a Von Mises distribution and calculating the probability based on integration, the result cannot be more than 1.0, which allows for an interpretation of the value as an estimation of the true probability.

Looking at the distribution in Figure 8.1 shows that the estimation is quite good at estimating when a person is not looking at the target. The method only wrongly estimates 9 values for test 2 thereby having 98.8% true negatives. The results for test 1 are okay with a result of 44.3% true positives. Combined, this results in 73.0% of the estimations being correct. This seems quite promising, though the estimation for test 1 could be improved.

When looking at the specific orientations and the method's performance, some of the cases stand out. Two of these cases are the right orientation with glasses at a distance of 170cm or 310cm from the camera. Looking at the video, these cases are actually quite understandable. Two frames from them are shown in Figure 8.3. For both cases it seems as though the subject's hair and glasses were blocking the view of the eyes from the camera. So, even if it was a human estimating the same video, they may not be able to tell whether the subject could see the target either.

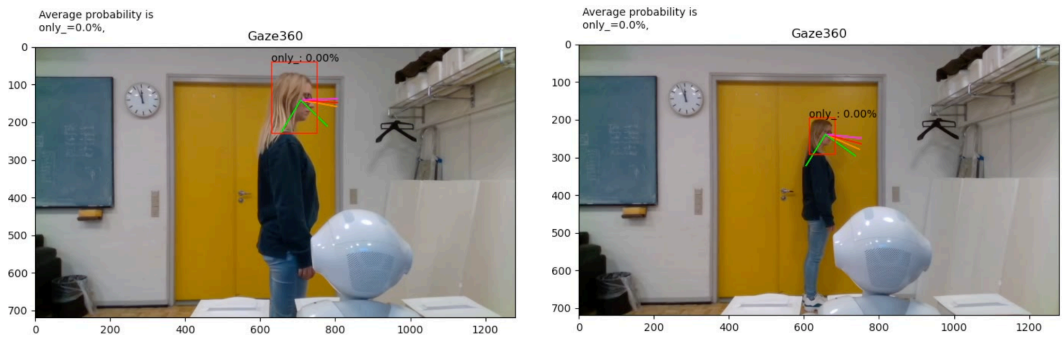


Figure 8.3: Two frames from test 1 where the subject wears glasses, at a distance of 170cm and 310cm from the camera and facing right. The algorithm run is method 3.

So does method 3 estimates whether a person is able to perceive a specific target? In truth, method 3 actually estimate the likelihood of the person fixating their gaze somewhere on the target. The work mentioned in Chapter 4 concludes that a person is able to perceive objects and movement as far as within their peripheral field of vision. So in order to imitate this theory, the method can be improved.

One of the frames which method 3 was not able to interpret correctly is seen in Figure 8.4. It seems as Gaze360 has such a high confidence that the Von Mises distribtion might be quite narrow, and thus result in a low probability within the targets.

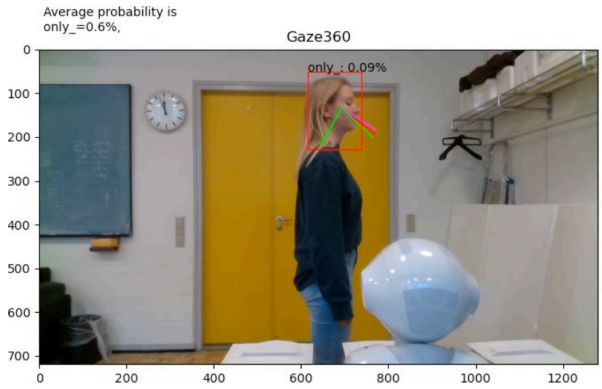


Figure 8.4: A frame from test 1, where the participant is facing forward, at a distance of 170cm from the camera, not wearing glasses. The estimation is using method 3.

Based on Chapter 4, it is known that the central vision is 2-5 degrees and the para-central is about 18 degrees. It can thereby be assumed that the participant is able to perceive a minimum of 18 degrees when applying overt attention. In Figure 8.4 the target is partially within 18 degrees of the gaze direction given by Gaze360. So can it be assumed that when a user is able to perceive only part of a target that the full target has been perceived? Or is a view of the full target required?

The question of how much of the target should be visible to the user is dependant on the scenario. If the target is complex with a lot of small details, the user might need to be able to see everything within the target. However, if the target is big cubes being rearranged in containers, the assumption of movements being perceived within the peripheral field of vision could to apply.

8.1.3.1 Method 3a

By extending the target's outer angles by 30 degrees to each side the method no longer focuses on estimating whether a person's gaze is fixated on the target. Now the method estimates whether the person's central field of vision is located so that part of the target is within the user's mid peripheral view ($30 - 60$ degrees on each side).

Now, looking at the same frame as Figure 8.4 but using method 3a in Figure 8.5. The estimation of the person was not above 0.5 using method 3, but for method 3a it is now estimated at 100%. As the participant actually was able to see the target in the frame, this is an improvement.

Overall, when comparing the distribution of method 3 and method 3a from Figure 8.1, method 3a looks to be able to estimate the data from test 1 with more confidence. The results from test 1 are 77.2% true positive and 86.6% true negative for test 2. Compared to method 3, the true positives for test 1 have improved quite a lot. Combined, the method estimates 82.1% of the probabilities as being correct based on the test data.

Based on the performance of the 3 methods, it seems as though method 3a has the most promising results given the test data, while method 2 has the worst. There are still many shortcomings for all the methods which are discussed in the rest of this chapter.

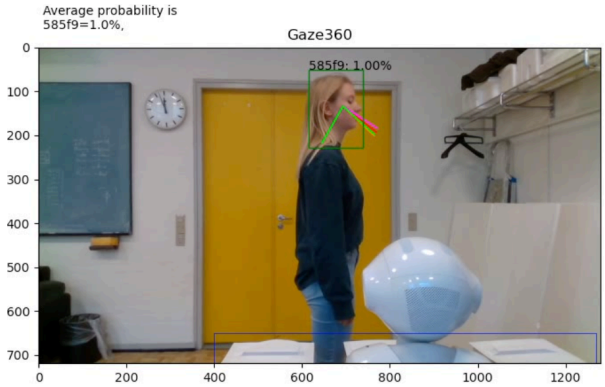


Figure 8.5: The same frame as from Figure 8.4, but using method 3a. The frame is from test 1 where the participant is facing forward, at a distance of 170cm from the camera, not wearing glasses.

8.1.4 Test data

The test is only conducted with 3 test participants, all at the same location. In addition, all test scenarios are only done once. This means that the results might not reflect the actual capabilities of the implementations. Ideally, the test scenarios should have been conducted with a variety of different conditions and with more diversity among the participants. Gaze360 is based on a quite diverse dataset. However, the power of this tool is not fully utilised in the experiment.

8.1.5 Cost

For the implementation to actually be used for human-robot interaction, the response must be within real-time computation. Currently the computational cost is very large, and the source of this is unknown. A large percentage of the computational cost is not due to the estimation by gaze360 nor the calculation of the distribution. It should be possible to optimize the implementation to run faster, but to do this the true source of the cause needs to be located.

8.2 Shortcomings

Overall, the basic concept of estimating an individual's perception of the world through a single target using gaze tracking is narrow. Chapter 3 discusses how peoples' senses are filtered by attention, which is then used to create people's inner perception of the world. Using only vision does not show the full image, as especially audio also has a large influence on our attention. Section 7.1.2 describes how people expect the robot to have a theory of mind similar to humans. The participants were actually quite confident in their assumption of the robot understanding changes of status only using sound. Even more confident than when the same test was conducted with vision. Thereby, simply knowing the accurate field of vision a human has, will not completely solve the issue. Even if the solution had 100% accuracy, there would still be missing interpretations from other senses.

The studies mentioned in Chapter 3 describe how specific tasks influence people's scanning patterns, and how selective attention may cause intentional blindness. Unless people are told specifically to keep an eye on the target, as they were for this experiment, it cannot be assumed that their attention is actually focused on their vision.

Further experiments are needed to fully understand whether people perceive scenes. I.e. what would a participant see, if they were told only to track the red cube and ignore the blue cube. Even though they look at the target, are they able to perceive where the blue cube is? These are all specific questions

related to how this experiment can be tested given the extent of what *to perceive the world* actually entails.

The estimation in the conducted experiment is 2 dimensional. However, gaze is actually 3 dimensional. So, if the target was above and behind the user, or a curtain was between the user and the target, all estimations would fail. The current methods do not have any depth perception, and do therefore not work in more complex cases.

8.2.1 Kappa angle

The last subject of discussion is the kappa angle, which was mentioned in Section 4.2.2. The kappa angle varies quite a bit between people with up to 4 degrees differences, which is almost the central field in itself. This is a large flaw for gaze tracking estimations which are not calibrated, as it proves impossible to estimate every person's gaze direction accurately. However, based on this study, the central field is only a small area within the peripheral view. The issue at hand is not whether a person's gaze is fixated on the target, but rather when the person is able to see and understand the change of events.

Having people need to calibrate every time they meet a social robot is not practical. The intended use of this study is to improve human-robot interaction. It would become bothersome to the users if they would need to calibrate every time they met a new robot. People are able to apply a theory of mind to a robot, so a robot having a false-belief is not unacceptable and might thereby be preferred to an accurate estimation.

8.3 Summary

Overall, method 3a showed the best results based on the test data with 82.1% of the estimations correct given a threshold of 0.5. In comparison, the worst results where produced by method 2 with only 59.3% being correct. In addition, when the test data was collected, the users were given specific instructions to look at the target. The assumption that people are able to perceive everything they see without instructions to focus on the task is naive. For the results to be accurate, multiple assumptions have been made, including that people are able to perceive rearrangements of cubes within their mid peripheral field of vision, that they use overt attention, and they are not inattentionally blind towards rearrangements.

CHAPTER 9
Further Improvements

Many improvements and new algorithms have been considered throughout the process of developing this work. Some of these improvements might not be large, but focuses on the immediate next steps for improving the implementation. Other ideas require more work and perhaps even starting from a clean slate.

9.1 Minor improvements

This section introduces four feasible improvements which would be simple next steps.

9.1.1 Real Time

One of the first improvements is analysing the large computational cost. The large computational cost makes it expensive to evaluate the solution, as a 4 second video could take about 5 minutes to analyse. Determining where the bottleneck is and how to remedy it would make testing much easier, and thus simpler to implement potential small improvements. The intention is also for the analysis to be in real time, as discussed in Chapter 8.

9.1.2 More Test Data

The second step would be to gather more test data. Some of the test data, as per the frames shown in Figure 8.3, are not optimal. By having the multiple participants conduct the same test, the results could be more informative and robust.

In addition, the estimations should also be done by a human. To simply ask a test person to look at a target is one thing, but this may force the algorithms to live up to an impossible standard, as e.g. the quality of the image might influence the level of precision. By having humans estimate the same test data, the implementation can be compared to the human results. The goal is to have the robot as reliable as a human in predicting what a person experiences.

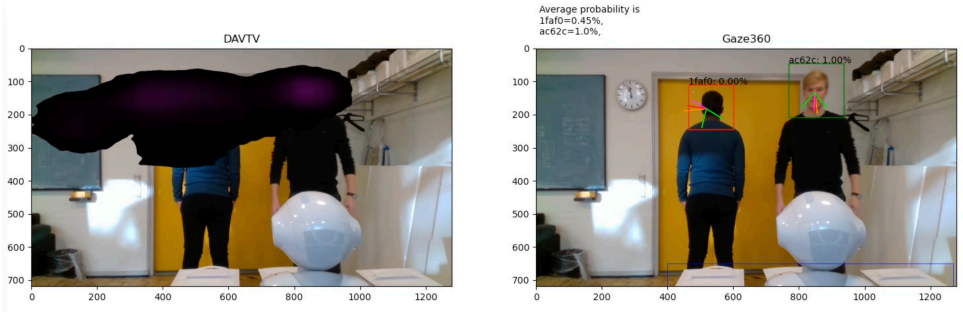
9.1.3 Estimating field of vision

The third improvement is experimenting with the extended angle, which was an addition in method 3a. The extension of 30 degrees was based on an assumption of visibility in the peripheral field of vision. It would be interesting to automatically calculate the optimal angular extension that results in the highest amount of correct estimations.

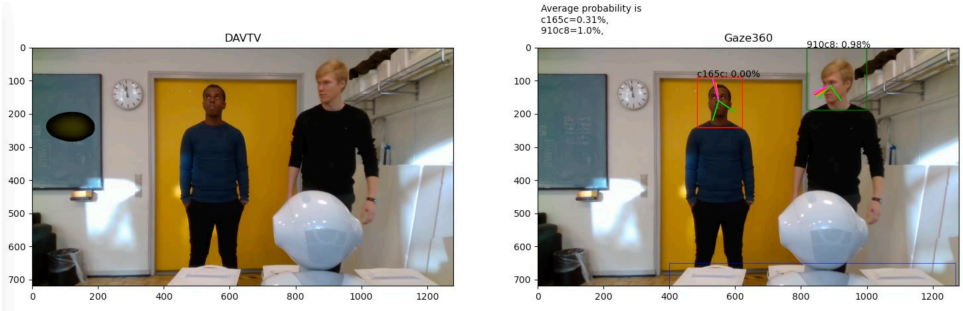
However, for this to be feasible, the implementation needs to be faster and more data is required. The model would need to calculate the optimal angular extension based on one dataset, while the evaluation would be calculated based on a different dataset. This is to generate an unbiased evaluation and thus avoid over-fitting.

9.1.4 Adding Additional Network

The final improvement is combining Gaze360 with the gaze tracking network *Detecting Attended Visual Targets in Video* (DAVTV). DAVTV is described in detail in Appendix E. Unlike Gaze360, DAVTV estimates a heatmap based on the estimated point of regard. It locates the target through a combination of a saliency map and gaze direction. The intent is to implement DAVTV as an extra probability estimation in addition to Gaze360. The probability of a person looking at the target object is greater if two separate solutions have the same estimation. However, the opposite is also the case, as shown in Figure 9.1.



(a) Methood 3a correctly estimate the target to be within the subject's field of vision. In contrast, DAVTV highlights the back wall incorrectly as the point of regard.



(b) DAVTV correctly estimates the blackboard to the left as the point of regard, while method 3a is not as precise.

Figure 9.1: Two examples showing how DAVTV and method 3a are able to estimate the gaze directions differently.

Figure 9.1 shows how the two neural networks might compliment each other, as one network can be correct, when the other is not. A technique for comparing the results in a reliable manner needs to be discovered, which is able to highlight each network's individual strengths. Based on a couple of videos using both networks (see Appendix D) it does not seem like one network is more correct than the other. Gaze360 might be a bit random at times and DAVTV only shows a result when it is sure that the point of regard is within the frame. The examples in Figure 9.1 are from the implementation of method 3a and illustrate how both networks are correct in different scenarios. In Figure 9.1a one of the participants is looking at the yellow door, so DAVTV correctly highlight it. However, the participant to the right is clearly looking at the target, which DAVTV do not highlight. In comparison, Figure 9.1b shows how method 3a estimates the participant at the right to look at the target, which he is not. While DAVTV does correctly estimate that the participant looks away.

9.2 Additional Experiments

These additional experiments are based on the knowledge gained within the process of creating the experiment. These changes are quite substantial, and might require starting from scratch.

9.2.1 Analyse the scene using 3D

One of the potentially large flaws is that the world is not 2 dimensional and people's field of vision is not circular, but extends more horizontally than vertically. The user might be looking directly towards the target, but if something is blocking the user's view, the camera would not be able to tell. E.g. if user 1 was moving the cube while blocking the view from user 2, user 2 would not be able to see where the cube was moved to, even if he or she was looking in the correct direction. As the experiment in Chapter 7 does not take the difference in horizontal and vertical angles into account when generating the estimations, these might potentially make a big difference.

So, to ensure higher accuracy and less assumptions, a solution should be developed in 3D. The output from gaze360 is spherical coordinates, which in turn are 3 dimensional. This would allow the solution to estimate the difference between looking down at the ground and looking straight ahead. The cameras used are actually also depth cameras, and are thereby able to detect the distance to objects within 2 m [63].

9.2.2 Sensor Fusion

The final improvement would be to implement sensor fusion. By creating a basic solution that includes an analysis of the participant's pose and the audio within the environment, the audio could simply determine whether a bottom-up salient sound might distract the participant to focus elsewhere. The robot does have a microphone, so this does seem like a possibility, but would require closer research.

In addition, a simple comparison of other information gained from the images could make a difference. Including head and hand pose estimation would create a more detailed understanding of the scene. Chapter 2 mentions how gaze tracking is not necessarily the only feature people base their understanding of each other on. Hand gestures and head movements also influence humans' interpretation of other people's world view. Hand pose estimation could determine whether a person is pointing, and head pose could allow more accurate estimations when the eyes are not visible.

The above-mentioned examples result in multiple different estimations based on different data. A technique called sensor fusion could then be used to compare results. This technique is crucial in e.g. autonomous vehicles. Sensor fusion uses the advantages and disadvantages within each source of input to conduct and create a combined and knowledgeable estimation. Autonomous vehicles therefore creates a perception of the environment the same way people do. Therefore, by utilising sensor fusion could greatly increase the accuracy of the solution.

9.3 Summary

Overall, four small improvements could be made to gain a greater result. These include reducing the computational cost, getting more test data, self-calculating the optimal extension of angle and comparing the result with a different source. The larger improvements are to implement the solution in 3D and / or to utilise more sensors than gaze tracking on which to base the estimation.

CHAPTER 10

Evaluation

The project plan and development is briefly discussed in this chapter as the thesis has changed throughout the process. The initial project plan is included in Figure 10.1 with more details in Appendix A. The emphasis initially was to create a plug-and-play product, which would preform estimations in real-time.

However, after the first couple of weeks it was evident that the topic was quite broad, and thus needed more research than expected. In addition, implementing Gaze360 required more work than expected as some of the functions used were deprecated and did not work on MacOS. About halfway through the process I contacted different specialists within gaze tracking and perception in order to gain a better understanding of my knowledge at the time. These meetings resulted in restructuring the project to focus on what perception is and how an algorithm can estimate it accurately.

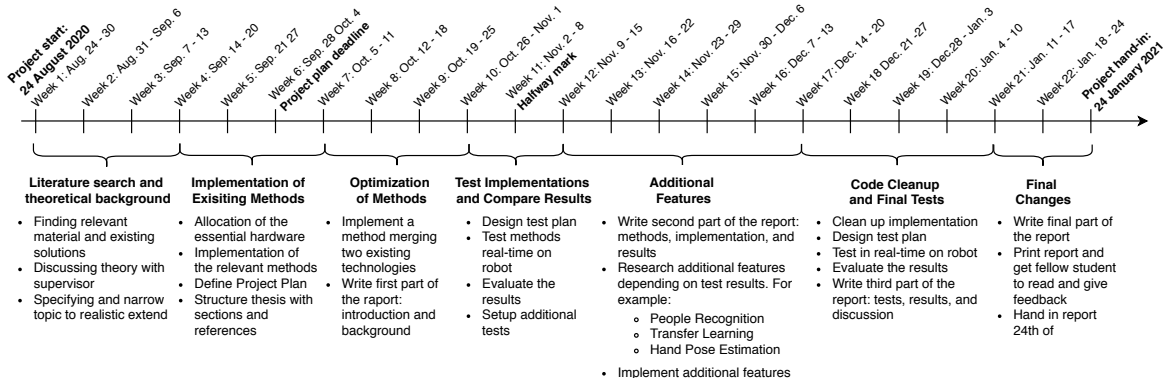


Figure 10.1: Overview of the process planned for the thesis.

Figure 10.2 shows how the actual process deviated from the original plan. The deadline was by extended a couple of weeks due to an unexpected delay in receiving student counselling assistance.

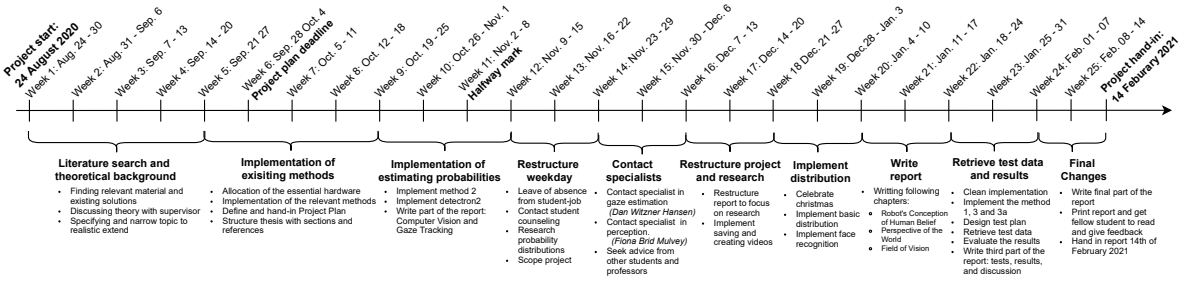


Figure 10.2: Overview of the actual process for this thesis.

Furthermore, this project has allowed me to personally learn how to work when demotivated and how

to structure days working from home, which was necessary due to the COVID-19 crisis. The importance of scoping the project and keeping to a timetable is highlighted, even if it means not knowing everything and simply accepting limited resources. Finally, one of the main skills developed throughout this project has been the ability to identify and reflect on technical scientific issues. Research within various fields of study created a greater understanding of the interaction between the different components within the issue.

If I was starting this project again, I would create a clearer initial scope and focus. The weekdays would have had a clearer structure by utilising tools such as the online study community, which is offered by the Student Counselling Service [64].

CHAPTER 11

Conclusion

People's cognitive perception of the world is complex as it is created based on people's senses guided by attention. Attention is the key component to understanding how humans generate individual perspectives as it is guided not only in a bottom-up fashion, but also based on top-down stimuli. Tasks, emotions, abnormalities, etc. all filter what people experience and notice. So, having peoples' perception of the world being estimated only through gaze tracking is a simplification, as perception is subjective.

People are able to estimate each other's perception and thus understand that they might be different. Theory of mind is an ability to distinguish and understand when a person has a false belief, which is an ability which is now being implemented in robots. However, for an implementation of theory of mind to work, it is essential to know when a person notices a change of status.

Though simple, using gaze tracking, it has been possible to estimate a person's ability to perceive a single target with a 82.1% accuracy. This accuracy was achieved by utilising the existing gaze tracking network Gaze360 in combination with detectron2's people detection and face recognition software. Then a Von Mises probability distribution could calculate an estimation of the probability of an individual having the target within their mid peripheral field of vision. The tests are based on a dataset where the participants were told to either be able to see the target or not see the target. It would be interesting to see how the results would vary if people where not specifically told what to look at. So, this thesis is a solid foundation for discovering how perception of the world is created and an inspiration for further work, which does seem to have great potential.

APPENDIX A

Project Plan

Introduction

Research utilizing artificial intelligence to learn social abilities processed by human have evolved radically during the last decayed. A significant understanding in social skills is the knowledge of every person having their own perception of the world. This project focus on using face recognition and gaze estimation to gain an understanding of what individuals see and estimate what they know.

This ability can create a more smooth interaction between person and machine, as the machine will understand what each individual will need to be informed off and avoid redundant explanations. An example where such technology can be utilized is in a shop scenario. Say a costumer looks at several products, but leaves without buying anything. Then when the same customer returns the system will be able to recognize the person and know which products' status changes the customer might be interested in. For instance, if one of the products the customer was previously looking at are now on sale. The machine will now not simply inform the customer about all products on sale, but only mention the new additions which it knows the customer might be interested in.

Project Plan

A Project Plan, see in Figure A.1, is created to ensure the desired outcome of the project is reachable within the given time frame. The plan has taken vacations and other potential unforeseen conflicts into consideration. The details about the process is described in Subsection A.0.1.

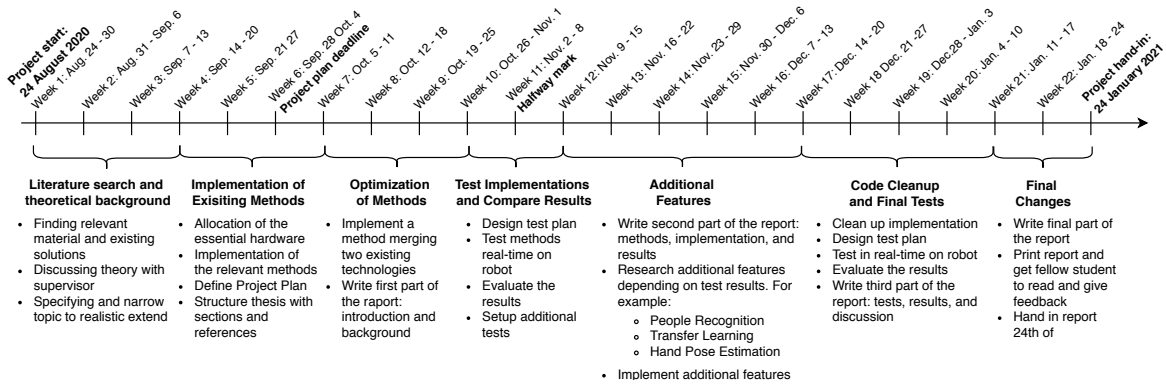


Figure A.1: Overview of the process planned for the thesis.

The timeline is a guideline and not the exact process of completing the thesis. Modifications to the plan will occur as the process will be developed using agile development adding new features in several iterations.

A.0.1 Details

Throughout the project ad hoc meetings with the supervisor will be held to discuss potential questions and dilemmas. These will be planed according to the need so that the development can keep as close to the plan as possible.

As seen in Figure A.1, there is a set of sub-goals to complete the written report throughout the project plan. This will benefit the final report by ensuring that the material is written with a fresh perspective and thereby avoid having the report miss crucial considerations.

APPENDIX B

Test Plan

This plan is designed to test the accuracy of the suggested solutions. The tests will all be recorded and analysed afterwards using the solution. The setup for all tests will be identical and similar to the setup created by Dissing et al. mentioned in chapter 2, [8].

Test: One person

This test is to understand how accurate the gaze tracking solution is to estimate whether an individual is looking at an object or not. The object will be the same for all targets. In the first test the participant will always be able to see the target object, while the second test the participant will never be able to see the target.

The following test will result in 6 recordings per. participant, which result a total of 93 different angles and gaze directions.

Target is Visible

The participant must *always* be able to see the target while moving around the scene given the parameters described in the following list.

- **Distance to Camera**: 310cm, 240cm and 170cm.
- **Body Pose**: Facing the left, front, right
- **Head Pose**: Facing front, up, left, down right (when turned to the right do not look away from the object and the same the other way).

An example of how to the participant is supposed to move around in the test is described in the following list.

1. The distance to camera is 1.7**m** and the participant's **body is facing left**.
2. The head moves up, down and right (always looking at the target)
3. The body turns to **face the target**.
4. The head moves up, left, down, right and strait (always looking at the target)
5. The body turns to **face right**.
6. The head moves up, down and left (always looking at the target)

The steps are repeated in new recordings for the remaining distances.

Target is **not** Visible

The participant must *never* be able to see the target, unlike the previous test. The method of testing is very similar, however, the participant must never be able to see the target. The parameters therefore change a bit, as described in the following list of parameters.

- **Distance to Camera**: 310cm, 240cm and 170cm.
- **Body Pose**: Facing the left, front, right and back.
- **Head Pose**: Look up, left, down, right and strait for all body poses, but never looking at the target.

This can test whether the solution is able to notice a difference from a person being able to see an action or object, and one not being able to see the target. This is essential for being able to utilize the solution for estimating people's perceptions of the world.

Test: Multiple people

The next test mimicking the false-belief tasks mention in Chapter 2, but with a focus on how the person with a false belief gets it. The tests will all be of first-order belief systems with similar steps. The steps are described in the following scenario between two subjects **A** and **B**.

1. A and B can both see the target.
2. A moves puts a green cube in container 1
3. A turns his back to the camera.
4. B moves the green cube to container 2.
5. A turns to face the camera gain.

The scenarios vary in step 3, given the following four cases:

1. A turns his **back** to the target and camera.
2. A turns his body to face **left** and look away from the target.
3. A turns his body to face **right** and look away from the target.
4. A looks away from the target, but the body is still **facing the target**.

The reason for testing a story scene is that the solution must track two people, and be able to know when both people are able to see the target, and when one of them is not. This can show how the solution is able to improve the current simplified solution implemented in the paper by Dissing et al. [8].

The test will result in 4 recordings per. participant-pair, with 8 of the same test is conducted, but the roles for **A** and **B** are swapped.

APPENDIX C

Results

Method 1: Facing the Camera

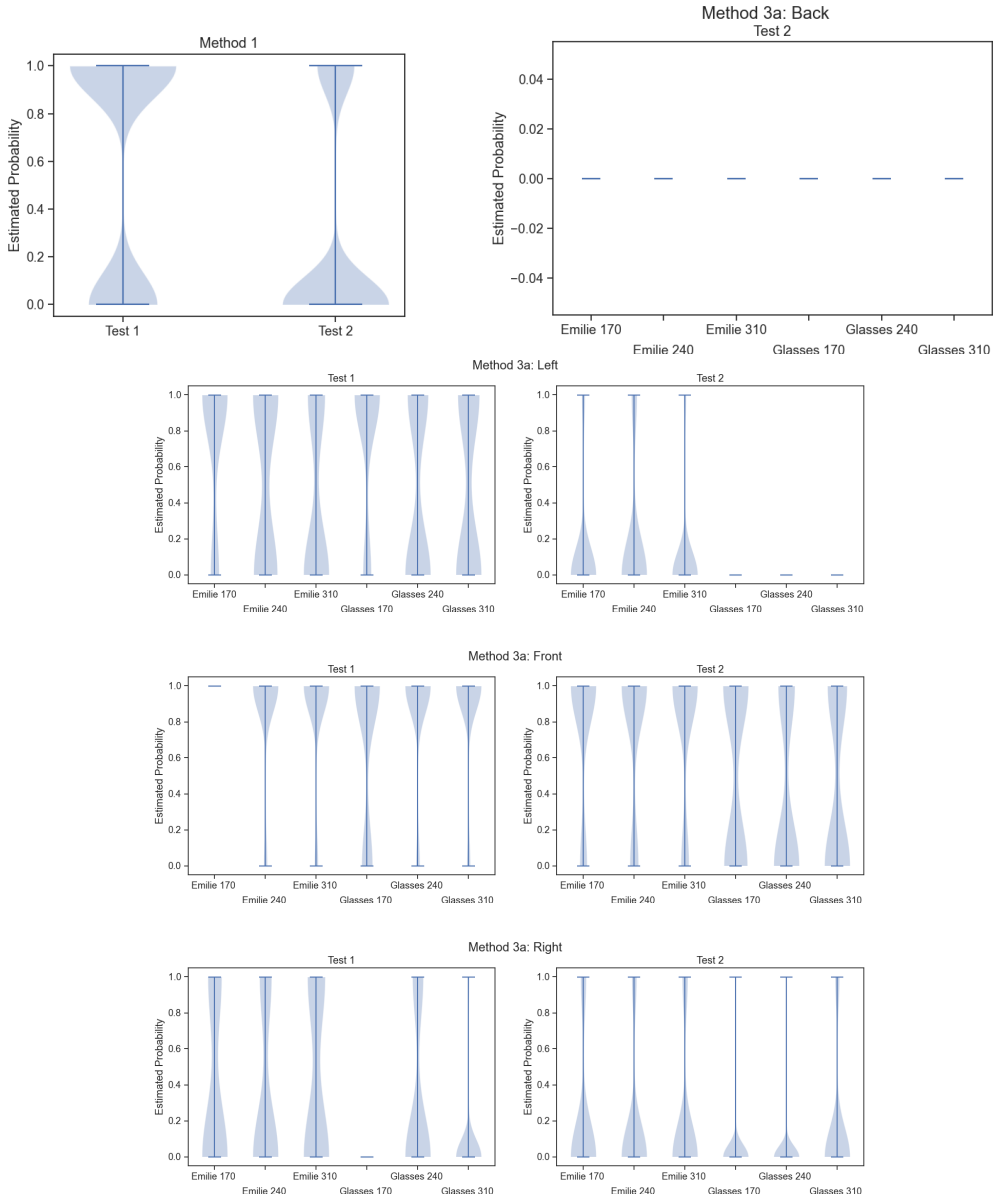


Figure C.1

Method 2: Mean gaze direction within target bounding box

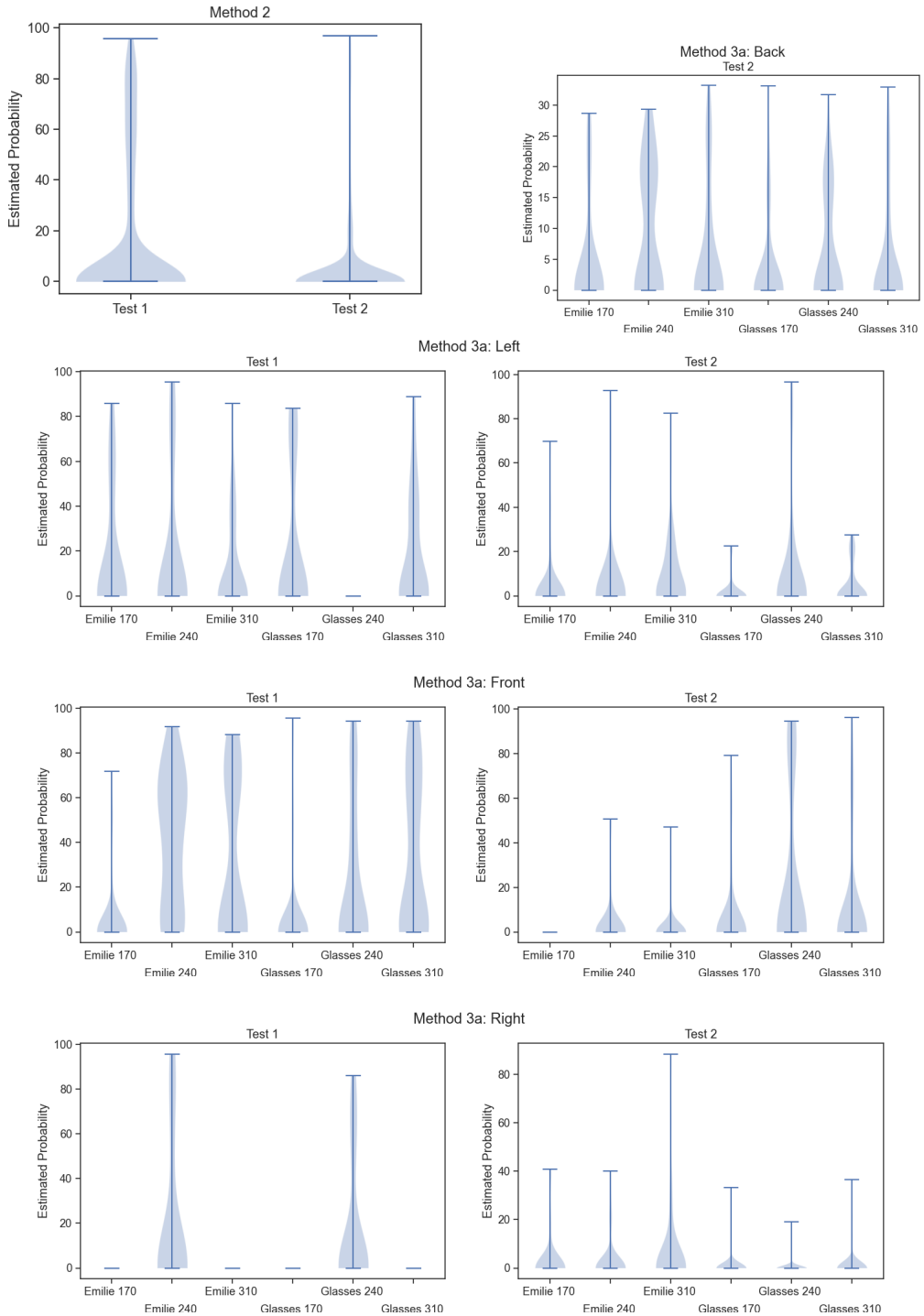


Figure C.2

Method 3: Von Mises distribution based on angles

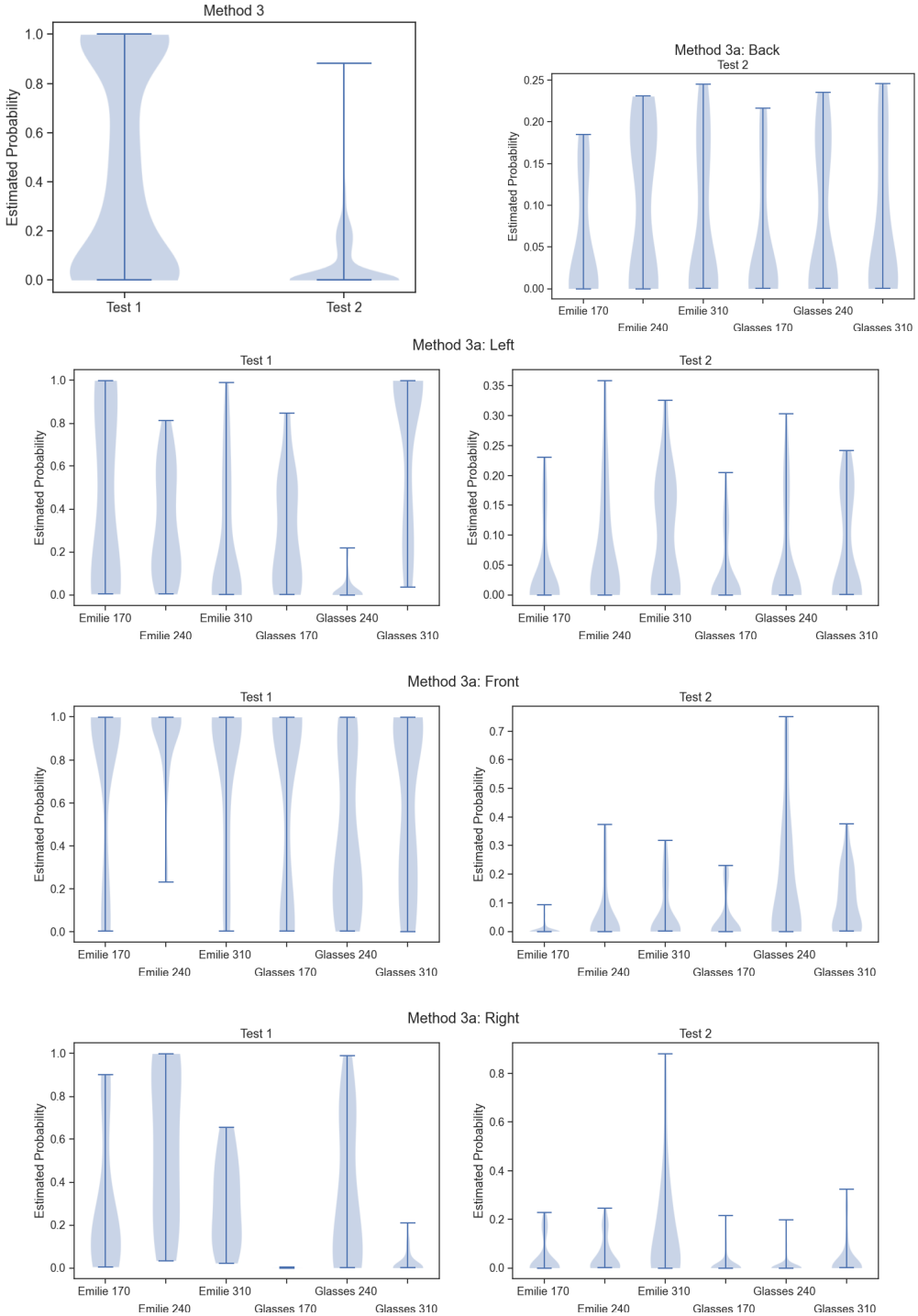


Figure C.3

Method 3a: Based on extended angles

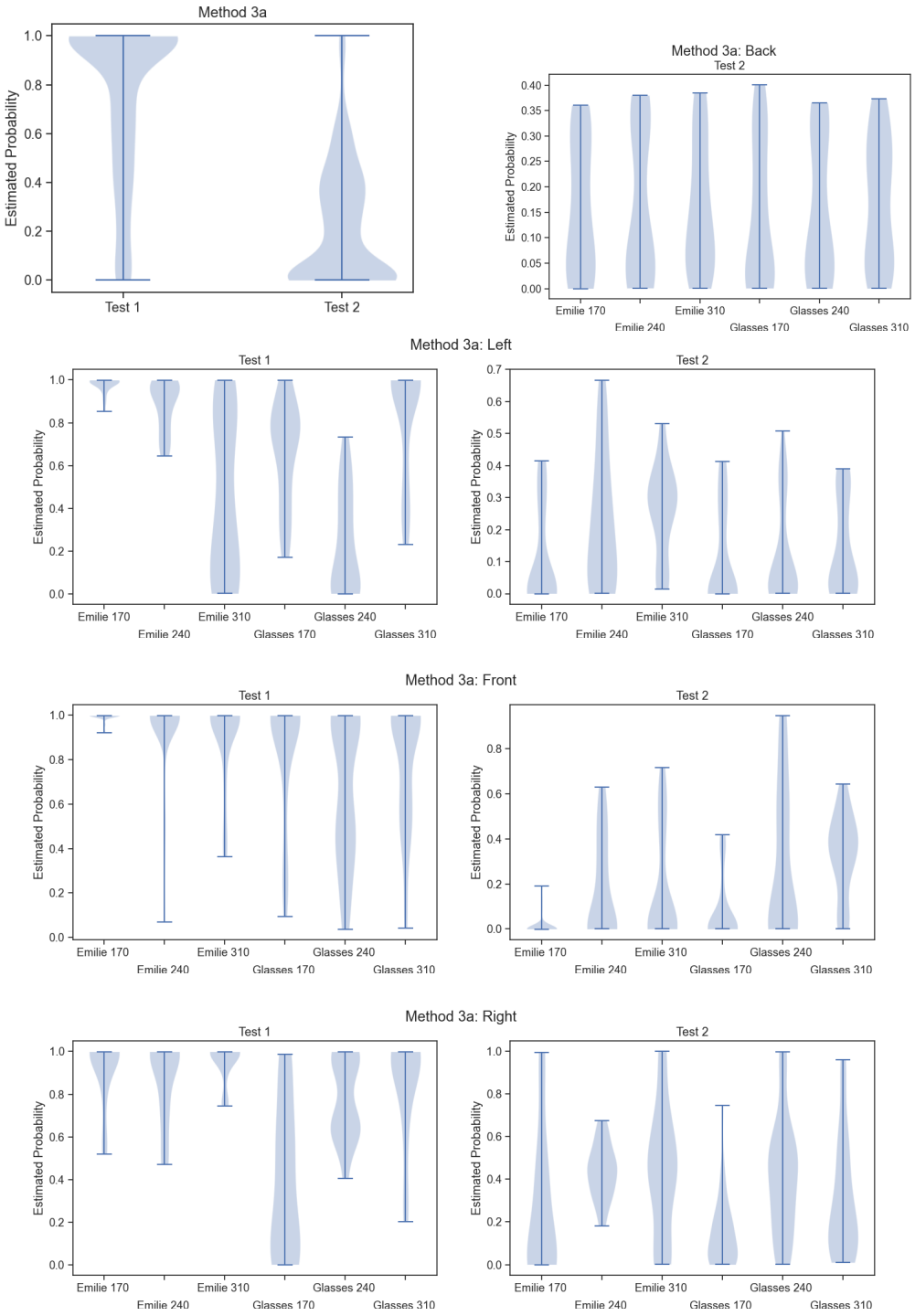


Figure C.4

Time

Method	Type	Mean	Median	Min	Max
Method 1	Total	315	296	149	782
Method 1	Gaze360	18	17	8, 4	45
Method 2	Total	323	299	148	777
Method 2	Gaze360	18	17	8.4	45
Method 3	Total	315	295	155	777
Method 3	Gaze360	18	17	8.6	44
Method 3	Distribution	0.21	0.19	0.11	0.48
Method 3a	Total	317	294	147	774
Method 3a	Gaze360	18	17	8.3	44
Method 3a	Distribution	0.22	0.21	0.11	0.5

Table C.1: Statistical results for time from analysis of 48 videos.

APPENDIX D

Github Code and Videos Demonstrating Solutions

All the implemented code, the test data and results are accessible through the following Github repository.

`https://github.com/s153762/Estimating_Peoples_Perception_of_the_World.git`

The raw and analysed videos are all uploaded in a folder called *Test_Data_Result*. The figures shown in this report is created by using the code in *Estimating_individuals_perspectives/Analyse_data/analyse_data.py*. The main code is run through the file *Estimating_individuals_perspectives/estimating_individuals_perspectives.py*.

For easy access one of the videos from Test 3 analysed using method 3a is uploaded to YouTube through the following link:

`https://youtu.be/94kip8cn4wk`

In addition, a video showing the initial errors with face recognition is shown through the following link:

`https://youtu.be/3AOQ5bDbKgI`

And finally, an implementation of DAVTV in comparison to method 3a from this solution is shown in the last link:

`https://youtu.be/QcPwwhqfaDg`

APPENDIX E

Detecting Attended Visual Targets in Video

Another interesting paper, *Detecting Attended Visual Targets in Video*, was mentioned in the survey described in Section 6.1.1 [42]. The network determines the point of regard of people in 2D videos and images, hence in a third-person view. This network is different as the focus is not only on detecting the point of regard, but also initially determined whether the gaze target is out of frame [65]. An example of the results of the network can be seen in Figure E.1. For simplicity, this network will be referred to as *DAVTV*.

Figure E.1: Results from [65]. Respectively the four images from left to right: The input, the first output of the deconvolution, the adjusted heatmap after modulation, and finally the prediction (yellow) and ground truth (red).

Like Gaze360, described in Section 7.1.1, a suitable dataset for training the network did not exist, resulting in the creation of the VideoAttentionTarget dataset [65]. The dataset contains 1331 video sequences of annotated data. It is constructed on various live interviews, sitcoms, reality shows etc. which are all available on YouTube. The length of a typical video sequence is $1 - 80$ seconds. The annotation of the data was created by selecting bounding boxes around each persons' head and the coordinates equivalent to gaze target. The dataset ends up having $33,4\%$ out of frame gaze annotations, and a total of 164.541 frames.

The network constructed in DAVTV is a bit complex, as it receives three different types of input and is constructed of three main parts. The architecture is illustrated in Figure E.2. The bottom branch seen in Figure E.2 is the *head conditioning branch*. This branch computes a feature map given the cropped head through a ResNet-50 network (head conv). This is in turn concatenated with the location of the head, represented in a binary image showing the area of the cropped head's original position. Finally, the concatenated head features are processed by a fully connected layer dubbed *attention layer*.

The *main branch* in Figure E.2 contains a scene feature map, which also have a ResNet-50 architecture. The input to the scene convolutional layer is a concatenation of the original image and the selected head's position. The scene feature map is then multiplied with the head feature map.

The last branch of the architecture is the *recurrent attention prediction module*. This branch encodes the output from the main branch using two convolutional layers. Then it process the encoding using a convolutional Long Short Term Memory network (CNN-LSTM), followed by four deconvolutional layers

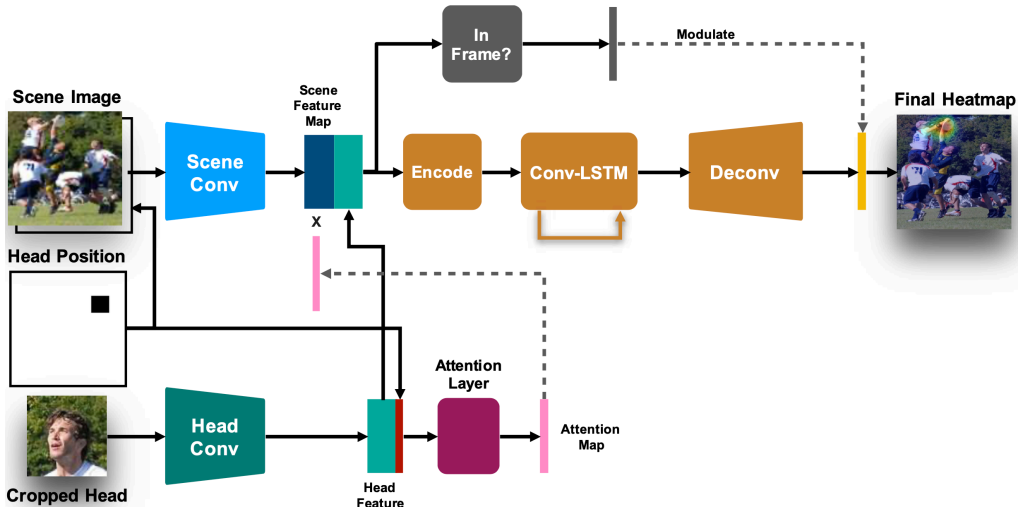


Figure E.2: The architecture of DAVTV model from [65].

which samples the features into a full-sized feature map. The feature map created by the deconvolutional layers can be seen as the second image from the left in Figure E.1. The last step is to modulate the feature map by a scalar which determines whether the attention is in frame. The scala is calculated through two convolutional layers and a fully connected layer, seen in the top of Figure E.2. The final heatmap is thereby the modulated map with a clipping to ensure that no value is less than 0.

Overall the evaluation of the DAVTV model shows state-of-the-art results on three datasets GazeFollow, VideoAttentionTarget, and VideoCoAtt. The preformance has an area under the curve of 0.921, which is close to the human performance of 0.924. It proves to be well suited to estimate the point of regard of the people in the image, even when the beholder might have a target outside of the scene visible in the image.

Bibliography

[1] Suzana Herculano-Houzel. *The human advantage: How our brains became remarkable*. 2016.

[2] Noriaki Mitsunaga et al. "What makes people accept a robot in a social environment-discussion from six-week study in an office." In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2008, pages 3336–3343. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4650785.

[3] Wikipedia contributors. *Human factors and ergonomics — Wikipedia, The Free Encyclopedia*. [Online; accessed 13-January-2021]. 2021. URL: https://en.wikipedia.org/w/index.php?title=Human_factors_and_ergonomics&oldid=997923831.

[4] Thomas B. Sheridan. "Human–Robot Interaction: Status and Challenges." In: *Human Factors* 58.4 (2016). PMID: 27098262, pages 525–532. DOI: 10.1177/0018720816644364. eprint: https://doi.org/10.1177/0018720816644364. URL: https://doi.org/10.1177/0018720816644364.

[5] Sean Andrist, Bilge Mutlu, and Adriana Tapus. "Look Like Me: Matching Robot Personality via Gaze to Increase Motivation." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pages 3603–3612. ISBN: 9781450331456. DOI: 10.1145/2702123.2702592. URL: https://doi.org/10.1145/2702123.2702592.

[6] Reuben M. Aronson et al. "Eye-Hand Behavior in Human-Robot Shared Manipulation." In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. Chicago, IL, USA: Association for Computing Machinery, 2018, pages 4–13. ISBN: 9781450349536. DOI: 10.1145/3171221.3171287. URL: https://doi.org/10.1145/3171221.3171287.

[7] H. B. Barua et al. "Let me join you! Real-time F-formation recognition by a socially aware robot." In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2020, pages 371–377. DOI: 10.1109/RO-MAN47096.2020.9223469.

[8] Lasse Dissing and Thomas Bolander. "Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic." In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2020. URL: http://www.imm.dtu.dk/~tobo/dissing2020implementing_proceedings.pdf.

[9] Heinz Wimmer and Josef Perner. "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception." In: *Cognition* 13.1 (1983), pages 103–128. URL: https://www.sciencedirect.com/science/article/pii/0010027783900045?casa_token=5bSl5N3P8nMAAAAA:zu0cPxTbkLcDkYI3MTwOCsvbtBCKnQbF1DZ8-hEFdcKEx84wUILah5yt51IP2aaU

[10] Chris Frith and Uta Frith. "Theory of mind." In: *Current biology* 15.17 (2005), R644–R645.

[11] Hans van Ditmarsch and Barteld Kooi. "Semantic results for ontic and epistemic change." In: *Logic and the foundations of game and decision theory (LOFT 7)* 3 (2008), pages 87–117.

[12] Stephen R.H. Langton, Roger J. Watt, and Vicki Bruce. "Do the eyes have it? Cues to the direction of social attention." In: *Trends in Cognitive Sciences* 4.2 (2000), pages 50–59. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(99)01436-9. URL: http://www.sciencedirect.com/science/article/pii/S1364661399014369.

[13] Grit Hein and Robert T. Knight. "Superior Temporal Sulcus—It's My Area: Or Is It?" In: *Journal of Cognitive Neuroscience* 20.12 (2008). PMID: 18457502, pages 2125–2136. DOI: 10.1162/jocn.2008.20148. eprint: https://doi.org/10.1162/jocn.2008.20148. URL: https://doi.org/10.1162/jocn.2008.20148.

[14] Trafton Drew, Melissa L.-H. Võ, and Jeremy M. Wolfe. "The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Observers." In: *Psychological Science* 24.9 (2013). PMID: 23863753, pages 1848–1853. DOI: 10.1177/0956797613479386. eprint: https://doi.org/10.1177/0956797613479386. URL: https://doi.org/10.1177/0956797613479386.

[15] Matei Mancas et al. *From Human Attention to Computational Attention*. Volume 2. Springer, 2016.

[16] Laurent Itti and Christof Koch. "Computational modelling of visual attention." In: *Nature reviews neuroscience* 2.3 (March 2001), pages 194–203. DOI: 10.1038/35058500.

[17] Fabian Hutmacher. "Why Is There So Much More Research on Vision Than on Any Other Sensory Modality?" In: *Frontiers in Psychology* 10 (2019), page 2246. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.02246. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.02246.

[18] Benjamin W Tatler et al. "Yarbus, Eye Movements, and Vision." In: *i-Perception* 1.1 (2010). PMID: 23396904, pages 7–27. DOI: 10.1068/i0382. eprint: https://doi.org/10.1068/i0382. URL: https://doi.org/10.1068/i0382.

[19] Jeremy M Wolfe. "Visual search." eng. In: *Current biology* 20.8 (2010), R346–R349. ISSN: 0960-9822.

[20] Rose M. Spielman et al. *Psychology*. December 2014. URL: https://openstax.org/books/psychology/pages/5-3-vision (visited on January 18, 2021).

[21] Steven M. LaValle. "Chapter 5: The Physiology of Human Vision." In: *Virtual Reality*. 2019, pages 127–153. URL: http://vr.cs.uiuc.edu/.

[22] MD Linda Lawrence and JD Moonika Jones. *Vision After Hemispherectomy, TPO Disconnection, and Occipital Lobectomy: An Introductory Guide*. November 2019. URL: https://www.brainrecoveryproject.org/wp-content/uploads/2017/06/FINAL-Vision-After-Hemispherectomy-TPO-Disconnection-and-Occipital-Lobectomy.pdf (visited on January 18, 2021).

[23] Matei Mancas and Olivier Le Meur. "Applications of saliency models." In: *From Human Attention to Computational Attention*. Springer, 2016, pages 331–377.

[24] *Width of Field of View: A fundamental characteristic of our visual system*. CogniFit. URL: https://www.cognifit.com/science/cognitive-skills/width-field-view.

[25] Mariani et al. *PART XIII: FACTS AND FIGURES CONCERNING THE HUMAN RETINA BY HELGA KOLB*. Stanford University School of Engineering. 1984. URL: https://webvision.med.utah.edu/book/part-xiii-facts-and-figures-concerning-the-human-retina/.

[26] David Heeger. *Perception Lecture Notes: LGN and V1*. New York University: Center for Neural Science. URL: https://www.cns.nyu.edu/~david/courses/perception/lecturenotes/V1/lgn-V1.html.

[27] Benjamin Thompson et al. "Peripheral vision: Good for biological motion, bad for signal noise segregation?" In: *Journal of Vision* 7.10 (2007), pages 12–12. DOI: https://doi.org/10.1167/7.10.12.

[28] Benjamin Tatler, Dan Hansen, and Jeff Pelz. "Eye Movement Recordings in Natural Settings." In: October 2019, pages 549–592. ISBN: 978-3-030-20083-1. DOI: 10.1007/978-3-030-20085-5_13.

[29] Dan Hansen and Qiang Ji. "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze." In: *IEEE transactions on pattern analysis and machine intelligence* 32 (March 2010), pages 478–500. DOI: 10.1109/TPAMI.2009.30.

[30] Abrahão Rocha Lucena et al. "Visual angles: Article on the importance of Multifocal Intraocular Lenses Implantation." In: *Rev Bras Oftalmol* 77.5 (October 2018), pages 268–271. ISSN: 0034-7280. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-72802018000500268&nrm=iso.

[31] Md Zahangir Alom et al. *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*. 2018. arXiv: 1803.01164 [cs.CV]. URL: https://arxiv.org/abs/1803.01164.

[32] Yann LeCun et al. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pages 2278–2324. URL: http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*. 2012, pages 1097–1105. URL: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[34] Justin Johnson. *Lecture 10 | Recurrent Neural Networks*. Stanford University School of Engineering. 2017. URL: https://www.youtube.com/watch?v=6niqTuYFZLQ.

[35] Afshine Amidi and Shervine Amidi. *VIP Cheatsheet: Recurrent Neural Networks*. 2018. URL: https://www.kaggle.com/blobs/download/forum-message-attachment-files/16155/recurrent%20neural%20network.pdf.

[36] Hongyuan Zhu et al. "XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, pages 2837–2846. ISBN: 9781450355520. DOI: 10.1145/3219819.3220105. URL: https://doi.org/10.1145/3219819.3220105.

[37] S. Seo et al. "Comparative Study of Deep Learning-Based Sentiment Classification." In: *IEEE Access* 8 (2020), pages 6861–6875. DOI: 10.1109/ACCESS.2019.2963426.

[38] Hien T. Nguyen and Thuan Quoc Nguyen. "A Short Review on Deep Learning for Entity Recognition." In: *Future Data and Security Engineering*. Edited by Tran Khanh Dang et al. Cham: Springer International Publishing, 2018, pages 261–272. ISBN: 978-3-030-03192-3.

[39] B Carpenter. "Coding chunkers as taggers: Io, bio, bmewo, and bmewo+." In: *LingPipe Blog. Available at: lingpipe-blog. com/2009/10/14* (2009).

[40] Yining Wang et al. "A compact and language-sensitive multilingual translation method." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pages 1213–1223. URL: https://www.aclweb.org/anthology/P19-1117.pdf.

[41] Ming Liang and Xiaolin Hu. "Recurrent convolutional neural network for object recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pages 3367–3375. URL: https://openaccess.thecvf.com/content_cvpr_2015/papers/Liang_Recurrent_Convolutional_Neural_2015_CVPR_paper.pdf.

[42] Dario Cazzato et al. "When I Look into Your Eyes: A Survey on Computer Vision Contributions for Human Gaze Estimation and Tracking." In: *Sensors* 20.13 (2020), page 3739.

[43] M. Al-Naser et al. "OGaze: Gaze Prediction in Egocentric Videos for Attentional Object Selection." In: *2019 Digital Image Computing: Techniques and Applications (DICTA)*. 2019, pages 1–8. DOI: 10.1109/DICTA47822.2019.8945893. URL: https://ieeexplore.ieee.org/abstract/document/8945893.

[44] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 7263–7271.

[45] Jifeng Dai. *Deformable Convolutional Networks*. ComputerVisionFoundation Videos. 2017. URL: https:https://www.youtube.com/watch?v=HRLMSrxw2To.

[46] Alireza Fathi, Yin Li, and James M Rehg. "Learning to recognize daily actions using gaze." In: *European Conference on Computer Vision*. Springer. 2012, pages 314–327.

[47] Mengmi Zhang et al. "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 4372–4381.

[48] Matthias Kummerer et al. "Understanding Low- and High-Level Contributions to Fixation Prediction." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. October 2017. URL: https://openaccess.thecvf.com/content_iccv_2017/html/Kummerer_Understanding_Low-_and_ICCV_2017_paper.html.

[49] M. Cornia et al. "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model." In: *IEEE Transactions on Image Processing* 27.10 (2018), pages 5142–5154. DOI: 10.1109/TIP.2018.2851672.

[50] Macario O. Cordel II et al. "Emotion-Aware Human Attention Prediction." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019. URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Cordel_Emotion-Aware_Human_Attention_Prediction_CVPR_2019_paper.html.

[51] Omer Sumer et al. "Attention Flow: End-to-End Joint Attention Estimation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. March 2020. URL: https://openaccess.thecvf.com/content_WACV_2020/html/Sumer_Attention_Flow_End-to-End_Joint_Attention_Estimation_WACV_2020_paper.html.

[52] Lifeng Fan et al. "Inferring Shared Attention in Social Scene Videos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[53] Xiaolong Zhou et al. "Two-eye model-based gaze estimation from a Kinect sensor." In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pages 1646–1653. URL: https://ieeexplore.ieee.org/abstract/document/7989194.

[54] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. "TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets." In: *Machine Vision and Applications* 28.5-6 (2017), pages 445–461. URL: https://link.springer.com/article/10.1007/s00138-017-0852-4.

[55] Cristina Palmero et al. *Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues*. 2018. arXiv: 1805.03064 [cs.CV]. URL: https://arxiv.org/abs/1805.03064.

[56] Petr Kellnhofer et al. "Gaze360: Physically Unconstrained Gaze Estimation in the Wild." In: *IEEE International Conference on Computer Vision (ICCV)*. October 2019. URL: http://gaze360.csail.mit.edu.

[57] Sam Thellman and Tom Ziemke. "Do You See what I See? Tracking the Perceptual Beliefs of Robots." In: *iScience* 23.10 (2020), page 101625. ISSN: 2589-0042. DOI: https://doi.org/10.1016/j.isci.2020.101625. URL: http://www.sciencedirect.com/science/article/pii/S2589004220308178.

[58] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[59] D King. "High quality face recognition with deep metric learning, 2017." In: *URL: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html (visited on 04/02/2021)* (). URL: https://github.com/ageitgey/face_recognition.

[60] Gary B. Huang et al. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical report 07-49. University of Massachusetts, Amherst, October 2007. URL: http://vis-www.cs.umass.edu/lfw/.

[61] Graham Upton and Ian Cook. *von Mises distribution*. 2008. DOI: 10.1093/acref/9780199541454.013.1717. URL: https://www.oxfordreference.com/view/10.1093/acref/9780199541454.001.0001/acref-9780199541454-e-1717.

[62] The Scipy community. *scipy.stats.mannwhitneyu*. 2016. URL: https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.stats.mannwhitneyu.html.

[63] *Depth Camera D435i*. Technical report. Intel RealSense. URL: https://www.intelrealsense.com/depth-camera-d435i/.

[64] Studenterrådgivningen. *Online study community*. URL: https://srg.dk/en/f%C3%A5-hj%C3%A6lp/gruppetilbud/online-studief%C3%A6llesskab/.

[65] Eunji Chong et al. *Detecting Attended Visual Targets in Video*. 2020. arXiv: 2003.02501 [cs.CV].

$\sqrt{17}$
$\Omega \int \delta e^{i\pi} = -1$
{2.7182818284}
ϑφερτυθιοπσδφγηξκλ
χ^2
$\sum$