TÉCNICO LISBOA

Eye-to-Eye: Gaze detection as a proxy for medical doctor behavior during appointments

Ricardo Andrade Antão

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Inv. Plínio Moreno López
Prof. José Alberto Rosado dos Santos Victor

Examination Committee

Chairperson: Prof. João Manuel de Freitas Xavier
Supervisor(s): Inv. Plínio Moreno López
Member of the Committee: Inv. Athanasion Mantalaris

June 2022

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

To Inv. Plínio Moreno López, Dr. Daniel Ferreira, Raquel Santos and Nuno André Silva, for all the patience, feedback and understanding of the subject. Without it, all the problems and decisions faced would have been much harder.

To my parents, who have always supported me and helped me whenever necessary.

To all my friends, who have been with me in this 5 year journey through university.

Resumo

O olhar do médico durante uma consulta realiza um papel importante na satisfação do paciente. Estudos indicam uma correlação positiva entre a satisfação do paciente e o tempo que o médico passa a olhar para o paciente. Adicionalmente, o efeito das videoconsultas no olhar do médico ainda não foi estudado. O objetivo deste estudo é de quantificar o impacto das videoconsultas no olhar do médico, comparando, entre consultas presenciais e videoconsultas, a quantidade de tempo que o doutor olha para o paciente. A população do estudo consiste em 14 doutores divididos entre 4 especialidades médicas: Ginecologia/Obstetrícia (4), Neurologia (3), Endocrinologia (3) e Medicina Geral e Familiar (4). Uma *pipeline* de estimação de direção do olhar foi implementada no ambiente clínico. Cada doutor gravou 20 consultas presenciais e 20 videoconsultas, processadas pela *pipeline* de forma a obter a percentagem de tempo da consulta em que o médico olhou para o paciente. O teste estatístico Mann-Whitney U foi usado para comparar as duas distribuições de percentagens. No caso em que uma diferença estatisticamente significante ($p < 0.05$) existisse, o Cohen's *d* foi usado para calcular o tamanho do efeito. No geral, foi verificado que todos os doutores, com exceção de dois, apresentam diferenças estatisticamente significantes ou tendências para olhar mais para o paciente em videoconsultas. Em relação às especialidades médicas, apenas Ginecologia/Obstetrícia não apresentou tendências para olhar mais para o paciente em videoconsultas. Consequentemente, em três das quatro especialidades estudadas, foram identificadas tendências para os médicos olharem mais para o paciente em videoconsultas.

Palavras-chave: Estimação de direção do olhar, Direção do olhar do médico, Interação clínica, Análise de biomarcadores digitais, Relação doutor-paciente

Abstract

The gaze of the physician during consultations plays an important role in patient satisfaction, with studies indicating the positive correlation between the time the physician spends looking at the patient and patient satisfaction. Additionally, the effects of virtual consultations on the gaze behaviour of the doctor have yet to be studied. Therefore, this study aims to assess the impact of virtual consultations by comparing the amount of time the doctor spends looking at the patient between face-to-face and virtual consultations. The study population consisted of 14 doctors divided between 4 medical specialties: Gynaecology/Obstetrics (4), Neurology (3), Endocrinology (3) and General and Familiar Medicine (4). An appearance-based gaze estimation pipeline was implemented in the clinical setup. Each doctor recorded 20 face-to-face and 20 virtual consultations, which were processed by the pipeline to obtain the percentage of time the doctor looked at the patient during the consultation. For each doctor, the Mann-Whitney U test was used to compare the two distributions. If a statistically significant difference ($p < 0.05$) existed, then the Cohen's *d* was used to calculate size effect. Overall, we found that all doctors, except for two, presented statistically significant differences or tendencies to look more at the patient in virtual consultations. Within medical specialties, only one specialty, Gynaecology/Obstetrics, presented doctors with no differences or tendencies between consultation environments, meaning that three out of the four specialties involved presented clear tendencies to look more at the patient in virtual consultations when compared to face-to-face consultations.

Keywords: Gaze estimation, Physician gaze, Automatic labelling, Clinical Interaction, Digital Biomarker Analysis, Physician-patient relationship

Contents

List of Tables

List of Figures

Nomenclature

CNN	Convolutional Neural Network
MAML	Model-Agnostic Meta-Learning
MLP	Multilayer Perceptron
PoR	Point of Regard
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent

Chapter 1

Introduction

The physician-patient relationship is a crucial component of the effectiveness of any health care system. Communication is one of the main components in a good physician-patient relationship where communication during medical interviews plays one of the most prominent roles [18].

The importance of non-verbal communication has gathered more and more attention from research studies that analyze the relationship between patient satisfaction and physician behaviour. A general practitioner can conduct between 120,000 and 160,000 interviews during a 40-year career [5]. Good communication skills are essential to reach improved management of chronic diseases like diabetes, hypertension, and others. In addition, patients are more adherent to medical recommendations and healthy behavioural changes when they are more informed and involved in the decision-making. Several studies indicate that the non-verbal cues of the physician are one of the most critical aspects of physician-patient communication [19, 38, 20, 37, 32].

Communication in the physician-patient relationship is divided into verbal and non-verbal communication. Verbal communication is defined as communication behaviour with linguistic content [38]. This is classified according to the model described by Bird and Cohen-Cole model [6]. According to this model, we can classify verbal interactions into three key functions: data gathering to understand the patient (gathering information), development of a rapport and responding to the patient's emotions (developing therapeutic relationship), and patient education and behavioural management (decision making and management). Non-Verbal communication can be defined as communication behaviour without linguistic content and is typically distinguished by which part of the body is being used to express the behaviour. Face non-verbal behaviours include smiling, gazing, frowning, eyebrow-raising, and facial expressivity. Body non-verbal behaviour is expressed through posture or gestures. Vocal non-verbal behaviour includes loudness, voice pitch, monotony, and speech rate.

Among the non-verbal behaviour, the gaze direction is one of the most important cues to analyze in non-verbal communication. Studies have concluded that there is a positive correlation between patient satisfaction and the amount of eye contact between the physician and the patient [37]. The amount of time the physician is gazing at the patient and not at the patient's health records on the screen can improve the patient's perception and cognitive functioning [48].

The majority of the works on this topic use manual annotation systems to quantify non-verbal behaviours. This process is costly and laborious, which is not convenient or scalable. Recently, some methods for automated annotation systems have been proposed [17, 48]. However, they were designed for a very constrained environment which would not translate well to other consultation offices/setups, since these methods don't allow to change the camera position inside the consultation office needed when operating in multiple consultation offices. Nonetheless, they provide the first approaches to automatic classification of physician's gaze during medical appointments.

This introductory chapter starts by placing the thesis in the context of this research topic. Followed by a topic overview and the objectives, where the intent of the thesis is clarified. The chapter ends with an outline of the thesis.

1.1 Problem Statement

The amount of studies published concerning the impact of non-verbal communication in the physician-patient relationship [19, 38, 20, 37, 32] has been increasing in recent years, which shows the importance non-verbal behaviours have in patient satisfaction. In this work we will focus on the automated analysis of the gaze direction of the physician during the medical interview, classifying it according to whether the physician is looking at the patient or not. It will aim to support the automated analysis of primary care visits focusing on effective communication in the physician-patient relationship.

Additionally, the rise in the use of virtual consultations due to the COVID-19 pandemic provided an entirely different environment for the physician-patient relationship. Its impact in the physician-patient relationship is yet to be fully understood. A preliminary observation made by the Luz Saúde group during 2019 (in the pre-pandemic era) perceived a rise in the eye contact between the patient and the physician during video consultations compared to the face-to-face consultations. This perception needs to be quantified and objectified clearly and scientifically.

1.2 Objectives

The primary objectives are (i) to quantify doctor's gaze behaviour during the consultation, (ii) to evaluate the impact of the virtual consultations in doctor's gaze behaviour and (iii) to evaluate how the change in environment affects different medical specialities.

The main performance indicator used to achieve our objectives will be derived from the quantification of the physician's gaze during a consultation (face-to-face or video consultation). Physician's gaze will be classified according to whether the physician is looking at the patient or at the screen / other areas of the room. After classifying all the samples in a consultation, we will calculate the percentage of time the physician spent looking at each of the possible areas. The main metric we are going to be analyzing and comparing will be the percentage of consultation time spent looking at the patient, which we will denominate as *Patient Percentage* (*Patient%*).

1.3 Contributions

The addition of technology to the clinical setting provides its own set of challenges that are independent of the accuracy of the technology by itself in normal conditions. There are several variables at play during a consultation that will affect the implemented technology performance, which, in some cases, can lead to it being unusable in the clinical setting. One of the major difficulties was due to the pandemic, which made mandatory the use of masks during consultations. Since the current automatic classification approaches were not tailored for masked users, doctors were given specific masks that were transparent in the mouth and nose regions of the face.

Therefore in this work, we implement a gaze estimation pipeline capable of estimating and classifying the doctor's gaze during face-to-face and virtual consultations. The pipeline is able to be applied to several consultation offices, only needing a mirror-based extrinsic camera calibration routine to be performed.

1.4 Thesis Outline

Chapter 2 begins with an explanation on how neural networks work and how to estimate their parameters. With the concept of neural networks introduced, we explain two particular types of neural networks used in state-of-the-art gaze estimation methods, the Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Then we introduce the concepts of camera intrinsic and extrinsic parameters needed to perform the gaze estimation task. In the end, we move on to gaze estimation methods. First, we explain the different types of gaze estimation tasks. After this, we introduce the different approaches made to solve these gaze estimation tasks. Then we compare the performance of the different approaches.

Chapter 3 contains a description the experimental design and protocol used in the study. First, it explains the characteristics of the statistical study. Then it explains the procedure used for data acquisition, explaining the different consultation environments and the recording interface given to the doctors to record consultations. It also explains the gaze estimation pipeline used to extract doctor's gaze direction during a consultation. In the end, it explains how the doctor's gaze direction was classified and what were the statistical tests used to compare the distributions from the virtual and face-to-face consultations.

Chapter 4 presents all the results obtained divided in medical specialties. The results of each individual doctor are analyzed in order to assess the impact of the virtual consultation environment. In the end, it provides a discussion about the conclusions and insights taken from the overall results and study contributions.

Chapter 5 provides the final conclusions of the study along with the future works to support the study.

Chapter 2

Background and State of the art

This section provides an overview the different techniques previously used in human gaze estimation tasks. We start with a theoretical background on neural networks and camera intrinsic and extrinsic parameters used in state-of-the-art gaze estimation. Then we explain the different types of existing gaze estimation methods.

2.1 Neural Networks

A Neural Network, also called Artificial Neural Network, is a computing system loosely based on the structure of animals brains. A Neural Network can be defined as a conjunction of connected nodes, each node is called a *neuron* or *perceptron* and each connection between nodes is called an *edge*, each edge is weighted by a *weight*.

Neuron

The fundamental building block of a Neural Network is called a neuron. A single neuron possesses a set of ***m*** inputs, $x_1, x_2, ..., x_m$ and an output $\hat{y}$. Each input x_i is weighted by its corresponding weight w_i. An additional input $x_0 = 1$ is sometimes added to the neuron and it is denominated of *bias* or threshold. A neuron converts the set of inputs $x_1, x_2, ..., x_m$ into output $\hat{y}$ by applying the propagation function, which consists of two sequential operations. First, it performs the weighted sum of the all the inputs. Then, a non-linear activation function f is applied to the result of the sum. The output $\hat{y}$ is the result of these two operations done sequentially and is given as:

$$\hat{y} = f\,(\sum_{i=1}^{m} w_i\, x_i\ + w_0) \tag{2.1}$$

Essentially, propagating data forward through the network is just applying sequentially linear transformations (weighted sums) and non-linear transformations (activation functions) on the data. Connecting multiple neurons in layers enables the network to perform complex operations and, consequently, gain the ability to learn very complex relationships between the inputs. The process of feeding the input data and propagating it through the network is called forward propagation.

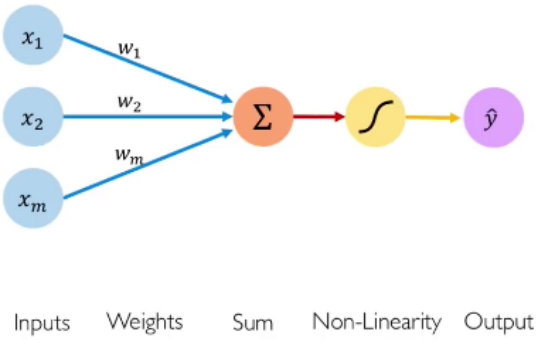


Figure 2.1: Diagram for a single neuron. [1]

Neurons give the ability to neural networks of performing two types of tasks: regression and classification. In neural networks, a regression task implies that the output of the network is a continuous value. It comes directly as the output of a neuron or set of neurons (depending on how many output variables there are). For example, a regression task can be estimating the gaze direction of a human. On the other hand, a classification task implies that the neural network classifies the input into a set of output classes. An example of a classification task would be the classification of a person on whether she is happy or sad according to its facial image, where happy and sad are both output classes.

Multilayer Perceptron

One of the simplest forms of Neural Networks are Multilayer Perceptrons (MLPs). In MLPs, the neurons are structured in layers where the input of neuron's in one layer is the output of neurons in the previous layer.

The first layer is called the input layer. Each neuron of the input layer has one input. The inputs of the neurons correspond to the input data for that task and vary from task to task. The middle layers are called the hidden layers and propagate the data forward through the network. The last layer is the output layer and each neuron in the output layer corresponds to one of the task's outputs.

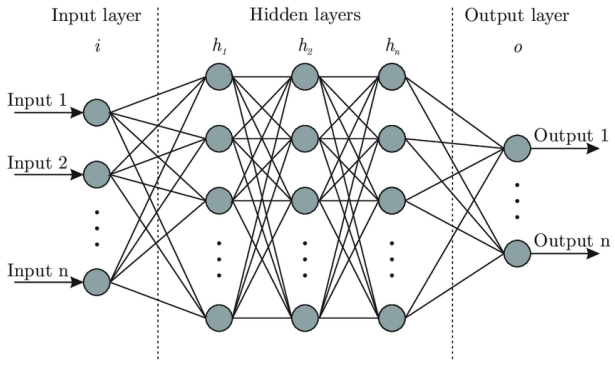


Figure 2.2: Multilayer Perceptron. [2]

[1] https://medium.com/analytics-vidhya/neural-network-part1-inside-a-single-neuron-fee5e44f1e[Date Accessed: 04-05-2021]

[2] https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051[Date Accessed: 04-05-2021]

2.1.1 Supervised Learning

Learning is the process of adapting the neural network to better handle the task at hand. Neural networks adapt to different tasks by changing the weights of the connections between their neurons. There are several learning paradigms that alter the way the network is trained. In other words, the process with which the weights in the network are changed. From the several paradigms: supervised learning, unsupervised learning, reinforcement learning and self-learning, we are going to focus on the supervised learning paradigm since it is the one used for training the majority of gaze estimation methods.

The supervised learning task is to teach a model how to yield the correct output from a certain input by using input-output pairs. This set of input-output pairs is the training dataset of the algorithm, where each training example consists of the input labelled with the desired output. This training dataset enables the model to learn over time how to perform a specific task correctly. The process of learning for a neural network consists in measuring its performance with a function, the loss function, and altering its weights so that the value of the loss function is minimized.

Loss function

The loss function measures the performance/accuracy of the model and it can be as simple as the mean squared error (MSE). The loss function is parameterized by the parameters of the model being trained. In the case of neural networks, the parameters of the loss function are the weights of the network. The result of the loss function is calculated after the forward propagation of the training data through the network.

Stochastic Gradient Descent

Essentially, the training of the model becomes an optimization problem of minimizing the loss function. This minimization problem is typically solved by applying an iterative method like the Stochastic Gradient Descent method. The Stochastic Gradient Descent (SGD) is an iterative method for minimization of an objective function $Q(w)$ where w are the parameters of the objective function. As the name indicates, the SGD makes use of the gradient of the objective function at each iteration. It takes advantage of the fact that the symmetric of the gradient of the function at any point always gives the steepest descent in the function. At each iteration t the parameters of the objective function are altered according to the following rule, where η is the learning rate of the algorithm:

$$w_t = w_{t-1} - \eta \nabla Q(w_{t-1}) \tag{2.2}$$

At each iteration i we move the value of the parameters w_i in the direction of the steepest descent, given by the symmetric of the gradient of the function. This way we minimize the value of the objective function in function of the parameters w.

The SGD method is used in Supervised Learning to minimize the loss function in function of the parameters of the model being trained. In neural networks, the objective function $Q(w)$ is the loss function

chosen, and the parameters w are the weights of the neural network. The training of a neural network involves applying the SGD to the loss function, obtained after the forward propagation of the training data through the network. Consequently, computation of the gradient of the loss function in relation to each weight of the network is needed. This computation is performed with the Backpropagation algorithm.

Backpropagation

In a neural network, to minimize the loss function, the network alters its weights. The alteration of the weights is done according to the SGD algorithm, calculating the gradient of the loss function in relation to the weights of the network and then updating the weights according to (2.2).

Backpropagation refers specifically to the algorithm used to calculate the gradients of the loss function. Due to the way neural networks operate, the gradient of the loss function in relation to each weight of the network needs to be calculated by propagating the loss backwards through the network with the application of the chain rule. The chain rule is used to calculate the derivative of composite functions, and it is given by:

$$\frac{\partial z}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x} \tag{2.3}$$

where $z = f(g(x))$. The output of the network $g(x)$ can be seen as a composite function of the form:

$$g(x) = W^{(L)} f(W^{(L-1)} \cdots f(W^{(1)} x)) \tag{2.4}$$

where L is the number of layers in the network

By propagating the gradient from layer $i+1$ to the previous layer i, the backpropagation algorithm is able to calculate the gradients in respect to every weight. The update of the weights is done according to the SGD method using (2.2) in the following way:

$$W_t^{(i)} = W_{t-1}^{(i)} - \eta \cdot \frac{\partial C(y, g(x))}{\partial W_{t-1}^{(i)}} \tag{2.5}$$

where t is the iteration number and i the layer number. Comparing with (2.2), the objective function $Q(w)$ corresponds to the loss function $C(y, g(x))$, where $g(x)$ is dependent on the weights of the network like in (2.4). The parameters w of the objective function correspond to the weights of the network $W^{(i)}$.

2.1.2 Convolutional Neural Networks (CNNs)

A Convolutional Neural Network (CNN) is a type of Neural Network. CNNs have been gaining popularity in recent years for Computer Vision tasks [1, 43]. CNNs consist of a set of convolutional and pooling layers followed by set of fully connected layers. There are various architectures of CNNs, some examples are LeNet [31], ResNet [21], AlexNet [30], among others. A CNN can be seen as a conjunction of 2 steps: Feature Extraction and Classification. The Feature Extraction step is performed by the convolution and pooling layers. During this step, CNNs take an input image and extract various characteristics (features)

from the image. Features can range from something as simple as lines and edges in the picture to more abstract (high-level) features that consist of combinations of simpler features. The Classification step is performed by the Fully Connected Layers, which are, in essence, an MLP.

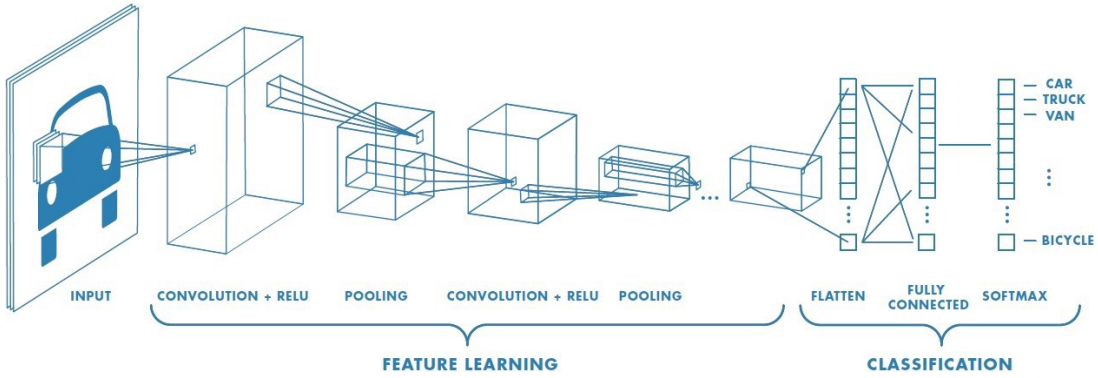


Figure 2.3: Convolutional Neural Networks. [3]

Convolutional Layer

The Convolutional layer gives the CNN the ability to capture spatial and temporal dependencies in the image through filters and the convolution operation. The filter is a conjunction of kernels, with different kernels being able to extract different features like edges or lines. The convolution operation extracts features from the input image according to the kernel used. Typically the first convolutional layers in the network extract low-level features such as edges, colour and gradient orientation. Using these features as inputs of other Convolutional layers, the CNN can extract high-level features, combining the previously obtained low-level features. The output of the Convolutional layers is sometimes called a feature map.

Pooling Layer

The Pooling layer is responsible for condensing the convoluted features returned by the Convolutional layer. The main objective of this operation is to reduce the spatial size of the feature maps, leading to the reduction of the computational power required in the processing of the data. Additionally, the pooling operation also extracts dominant features, which are rotational and positional invariant. Pooling works by passing the kernel through the image. There are two types of Pooling: Max Pooling and Average Pooling. Max Pooling returns the maximum value from the portion of the image covered by the kernel. On the other hand, Average Pooling returns the average of the values from the portion of the image covered by the kernel. The Pooling layer is responsible for increasing the efficiency of the CNNs training by reducing the dimensionality of the data and extracting the most dominant features simultaneously.

Fully Connected Layers

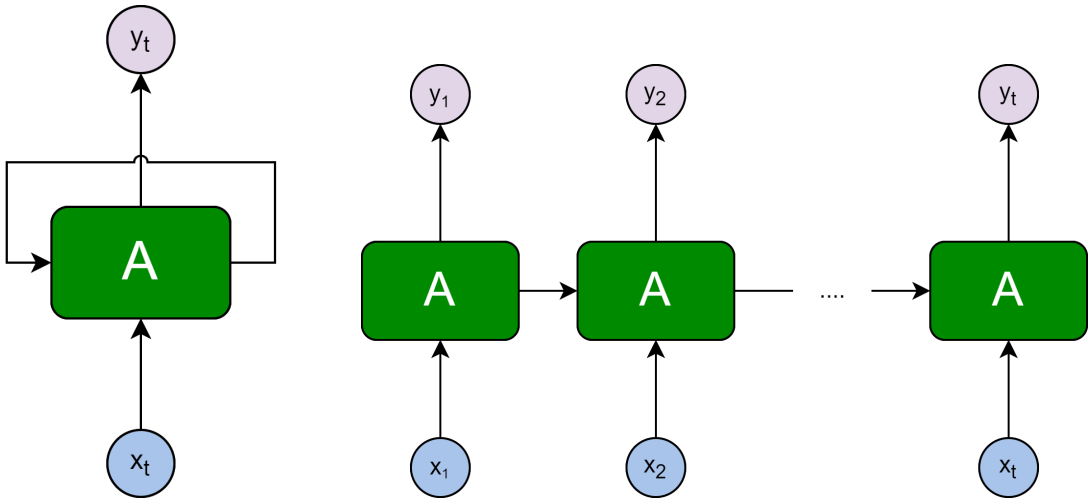
The Fully Connected Layers take the feature map returned by the convolutional and pooling layers and output the classification of the input image. This section of the CNN is an MLP that learns how to perform

[3] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [Date Accessed: 04-05-2021]

the wanted task with the feature map as its input.

2.1.3 Recurrent Neural Networks (RNNs)

A Recurrent Neural Network (RNN) is also a type of Neural Network like the CNN. A RNN is a type artificial neural network which deals with temporal/sequential data of variable length t, where input is a sequence $x = \{x_1, x_2, x_3..., x_t\}$ and output is also a sequence $y = \{y_1, y_2, y_3..., y_t\}$, the sequence is typically a temporal sequence although any type of sequence works in a RNN. Due its ability to model temporal dependencies, the RNN has been gaining popularity in the fields of natural language processing (NPL), speech recognition and image captioning where the current output is highly dependent on previous events and not just on the current input. To model time dependencies, RNNs are designed to have loops in the network, which allow the information of prior inputs to persist, by passing information from one step to the next, this is illustrated in Figure 2.4a



(a) RNN Diagram (b) Deconstructed RNN

Figure 2.4: RNN representations: The typical RNN can be represented as a single block with a loop or a sequence of identical blocks connected in a chain, one for each input

As a simplification, the RNN can be deconstructed into multiple copies of the same network, one for each input in the sequence, like in Figure 2.4b. This way, it is much easier to visualize how RNNs are able to create and model dependencies between sequential inputs making information about previous inputs persist in the networks, while in a typical CNN or MLP sequences of inputs and outputs are independent from each other. The ”A” blocks in Figure 2.4 represent the architecture of the RNN. There are several types of architectures of RNNs, however the most popular and widely used type is the Long-Short Term Memory (LSTM) [22].

2.1.4 Long-Short Term Memory

The Long-Short Term Memory first proposed in [22] had the objective of dealing with the limitation of conventional RNNs of not being able to model long-term dependencies. The key idea behind the LSTM is the cell state s_t, which keeps information about the previous events. The cell state value is controlled

through gates that control what information is added or removed from the cell state, as well as what information is outputted from the cell. In essence, a LSTM is composed of a cell and 3 gates: forget gate, input gate and output gate and it is illustrated in Figure 2.5.

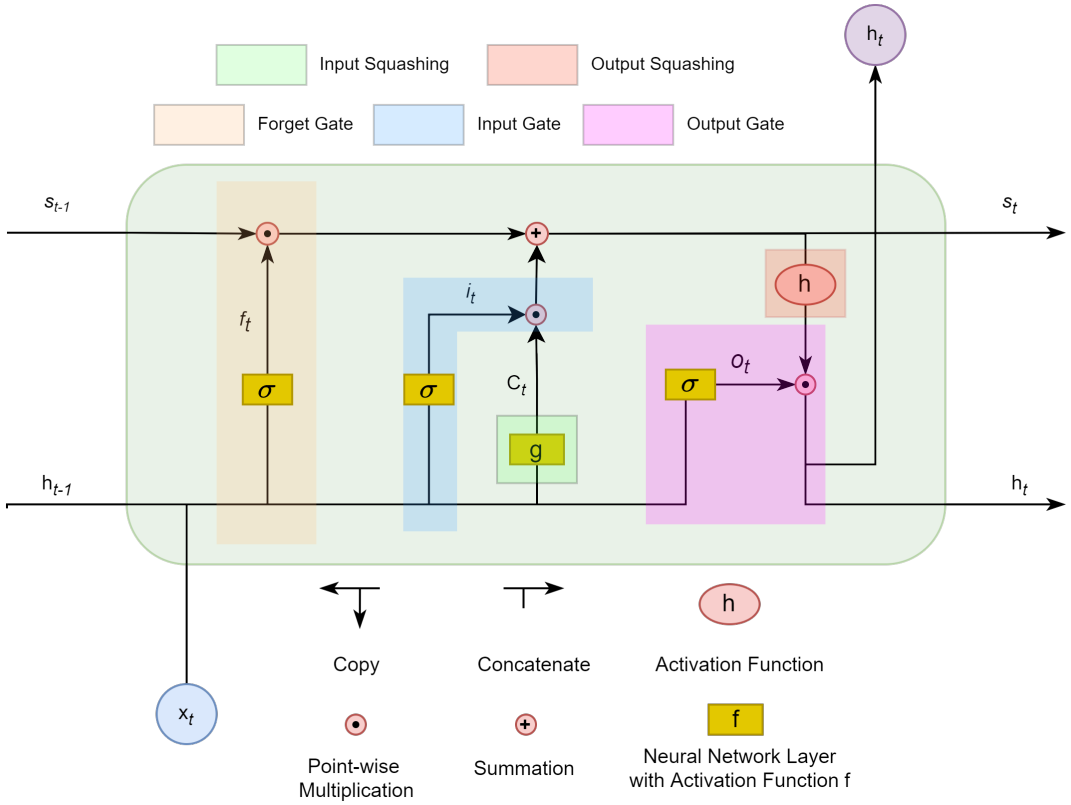


Figure 2.5: Long-Short Term Memory Module: The LSTM has 3 inputs (previous state, previous output and current input) and 2 outputs (state and current output). The process of turning the inputs into the outputs is carried out by the 3 gates and the 2 squashing operations performed on the data

The inputs of the LSTM are made up of the previous state s_{t-1} and the concatenation of the previous output h_{t-1} and the current input x_t. A gate, illustrated in Figure 2.6, is a way to control how much information is passed on to the next step. Each gate is composed of a neural network layer with a sigmoid activation function, similar to a hidden layer from Figure 2.2, followed by a point-wise multiplication. The

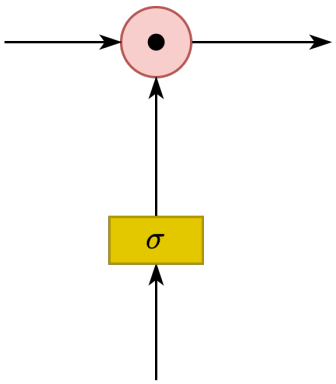


Figure 2.6: LSTM Gate

neural network layer has a neuron for each value of the input $[h_{t-1}, x_t]$ that works like the ones illustrated

in Figure 2.1, therefore its output is always between 0 and 1 (due to the activation function being a σ). When the output is 1 the gate keeps all the information, when the output is 0 the gate blocks / "forgets" the information. Each gate has a specific function in the LSTM, all related to controlling the information on the state of the cell s_t. the input gate controls what information from the inputs is stored in the state and the output gate controls what information is outputted by the cell as h_t. Alongside the gates, the LSTM also performs a *squashing* operation on the input $[h_{t-1}, x_t]$ and the state s_t. The objective of the *squashing* operation is to "squash" the values so that they are between -1 and 1, this helps to regulate the LSTM by reducing the size of the absolute value of the inputs. This is accomplished by passing the values through an activation function, normally the hyperbolic tangent $\tanh$. There are two different *squashing* operations: input *squashing* performed by the g layer and output *squashing* performed by the h function.

Forget Gate

The forget gate, shaded in orange in Figure 2.5, controls what information is removed / "forgot" from the previous state and it acts on the previous state s_{t-1}. The gate generates the vector f_t by passing $[h_{t-1}, x_t]$ through the activation layer. Equation (2.6) which is very similar to (2.1), describes the network layer operation, where W_f is the weight matrix of the network and b_f the *bias*.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.6}$$

The values of f_t define how much information from the previous state is forgotten by the LSTM.

Input Gate

The input gate, shaded in blue in Figure 2.5, controls what information from the input $[h_{t-1}, x_t]$ is stored in the state s_t. The input gate vector i_t is obtained in a similar fashion to the f_t vector from the forget gate, with (2.7). Additionally, before passing through the gate, the input is squashed to form the candidate values C_t, see (2.8).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.7}$$

$$C_t = g(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.8}$$

With the help of both the input gate and the forget gate, the state s_{t-1} can finally be updated through (2.9) with a linear transformation, obtaining the current state s_t.

$$s_t = f_t \cdot s_{t-1} + i_t \cdot C_t \tag{2.9}$$

The gates control how much information from the previous state is kept and how much information from the candidate values is added.

Output Gate

After obtaining the current state s_t, the output gate will decide how much information from the state should be outputted as h_t. For this, the gate generates vector o_t like the previous 2 gates, see (2.10).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.10}$$

Additionally, the state is *squashed* by passing it through the h function in the output squashing operation. After this, the squashed state is filtered by the output gate forming the output h_t, see (2.11).

$$h_t = o_t \cdot h(s_t) \tag{2.11}$$

2.2 Camera Parameters

A camera can be described by the pinhole camera model illustrated in Figure 2.7. The pinhole camera model describes the mathematical relationship between the points in the world coordinate system and their projection onto the camera image plane (Pixel Coordinate System). This mathematical relationship is parameterized by the camera's intrinsic and extrinsic parameters.

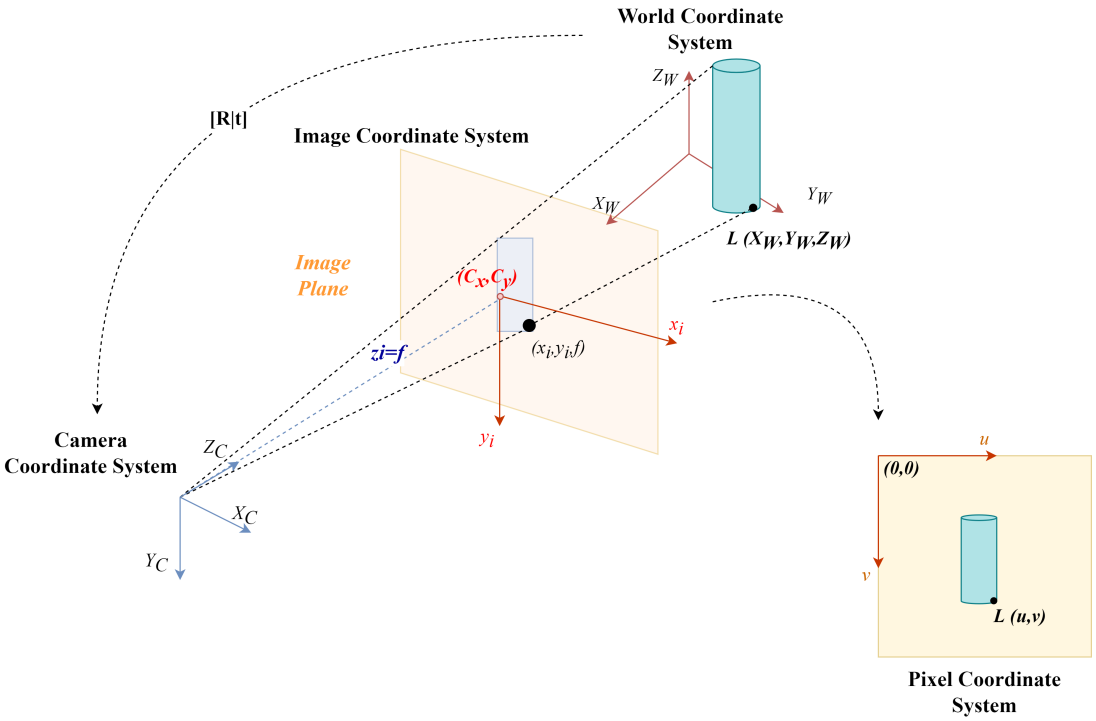


Figure 2.7: Pinhole Camera Model.

The intrinsic parameters of the camera describe its internal characteristics such as focal length, skew, distortion and image center. They are necessary to link the pixel coordinates of an image point with the corresponding coordinates in the camera coordinate system. Essentially, they describe the transformation of points in the Pixel Coordinate System to points in the Camera Coordinate System (CCS) and vice-versa.

The extrinsic parameters describe the camera's position in relation to the world coordinate system, defining the location and orientation of the camera reference frame in respect to the world reference frame. The extrinsic parameters describe the transformation of points in the world coordinate system to the camera coordinate system and vice-versa.

In gaze estimation for screen-based applications, the extrinsic parameters are used to convert 3D gaze predictions (vectors in the Camera Coordinate System) into 2D gaze predictions as *PoG*s (points of gaze) on a screen. The screen is described as a plane in the world coordinate system, typically referred as the Screen Coordinate System (SCS). Using the camera's extrinsic parameters, we can convert from CCS coordinates to SCS coordinates enabling the computation of intersections between 3D gaze directions and 2D screen planes in the CCS.

Therefore, we need to know both the intrinsic and extrinsic parameters of the camera. For this, we use two camera calibration routines. To find the intrinsic parameters of the camera, we use the routines offered by the *OpenCV* [8, 57, 7] library. To find the extrinsic parameters of the camera, we use a mirror-based calibration method from [47].

2.3 Gaze Estimation Methods

2.3.1 Gaze Estimation Methods Background

Gaze estimation objective is to estimate a subject's gaze direction. The earliest attempts at gaze estimation consisted in the detection of eye movement patterns like fixation, saccades and smooth pursuits [52]. These early methods attached sensors around the eyes to detect the movement patterns mentioned before. However the evolution of computer vision technology enabled the creation of modern eye tracking software devices, which captured eyes/face images of the subject from which gaze direction was inferred. In Figure 2.8 the history of gaze estimation methods is illustrated.

Modern gaze estimation methods powered by computer vision can be divided into three categories: 2D eye feature regression methods, 3D eye model recovery methods and appearance-based methods.

The first two methods consist of detecting geometric features of eye appearance such as contours, reflections and eye corners. 2D eye feature regression methods learn functions that map from geometric features from human gaze [25, 39]. 3D eye model recovery methods build subject-specific 3D eye models which are then fitted with the detected geometric features like infrared corneal reflections [15, 58], pupil center [50] and iris contours [2]. Both categories of methods need dedicated devices (infrared cameras or RGBD cameras) to detect geometric features. Additionally, 3D eye model recovery methods use subject-specific models, thus needing time consuming personal calibration sequences to estimate the subject specific parameters of the model.

Appearance-based gaze estimation methods learn functions that map directly from images to gaze direction by using image features like image pixels [34] or deep features [54]. In contrast to the two previous types of methods, appearance-based methods work with off-the-shelf web cameras. Regressing gaze directly from images is a difficult task due to the variety and complexity of eye appearance and

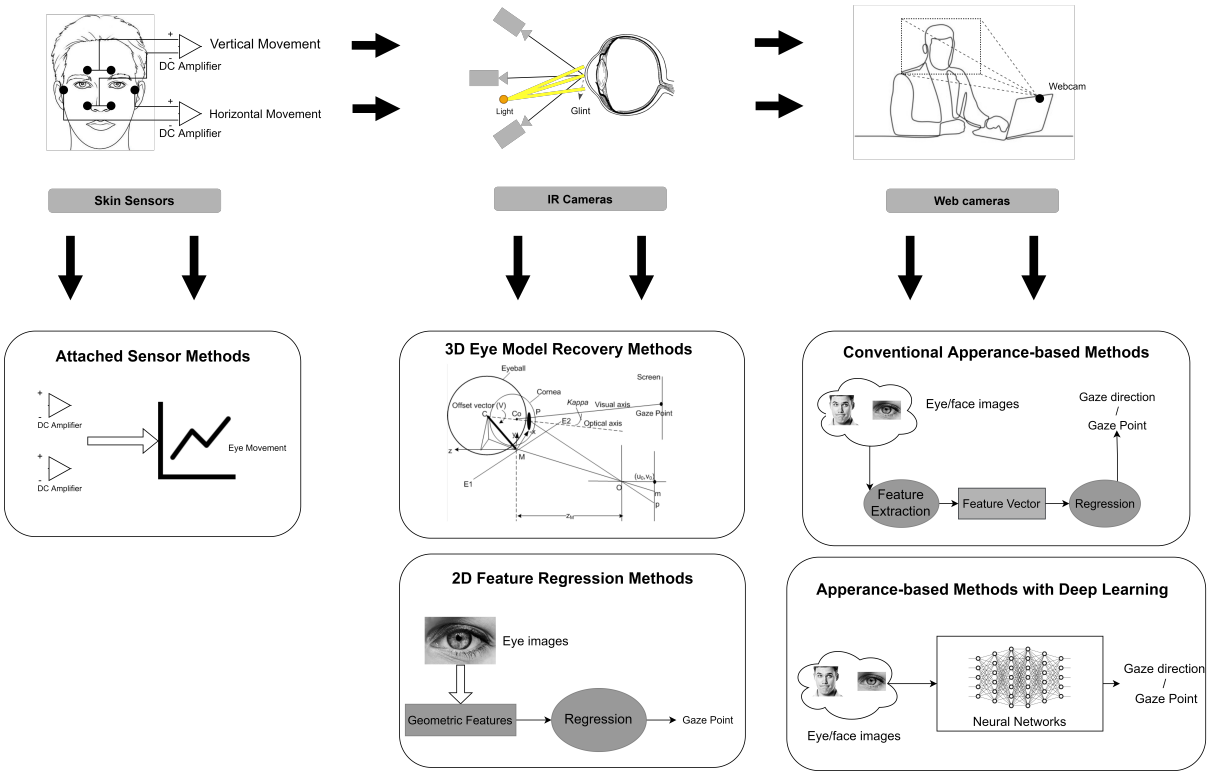


Figure 2.8: The Evolution of Gaze estimation devices and methods: Starting in the pioneering methods based around skin attached sensors. Then moving to the more robust but less accessible methods using IR and Depth cameras. In the end, the more recent appearance-based approaches using web cameras, that with deep learning have become both robust and accessible.

several models have been tested for this task *e.g.* neural networks [4], Gaussian process regression models [51], an adaptive linear regression model [34] and convolutional neural networks [54].

In this section we are going to focus on appearance-based methods due to being the chosen method for the study due to its flexibility, robustness methods and being the focus of the majority of the recent gaze estimation research [29, 11, 28, 13, 54, 56, 55, 23, 41, 42, 46].

2.3.2 Appearance-based Gaze Estimation Task

In the context of a appearance-based gaze estimation task, gaze should be understood as *gaze direction* or point of gaze (*PoG*). *Gaze direction* is represented as a 3D vector in the Camera Coordinate System and the *PoG* is represented as a 2D point in the target Screen Coordinate System. The typical setup (Figure 2.9) for gaze estimation tasks consists of a camera directed at the subject and a gaze target, normally a screen.

Therefore, the objective with gaze estimation tasks is to model the relationship between input images of a subject and its gaze direction/*PoG*. This divides gaze estimation into two different types depending on the output: 3D gaze estimation (gaze direction) and 2D gaze estimation (*PoG*). Additionally, gaze can be classified according to zones of the screen, this type of methods are classified as gaze zone estimation methods.

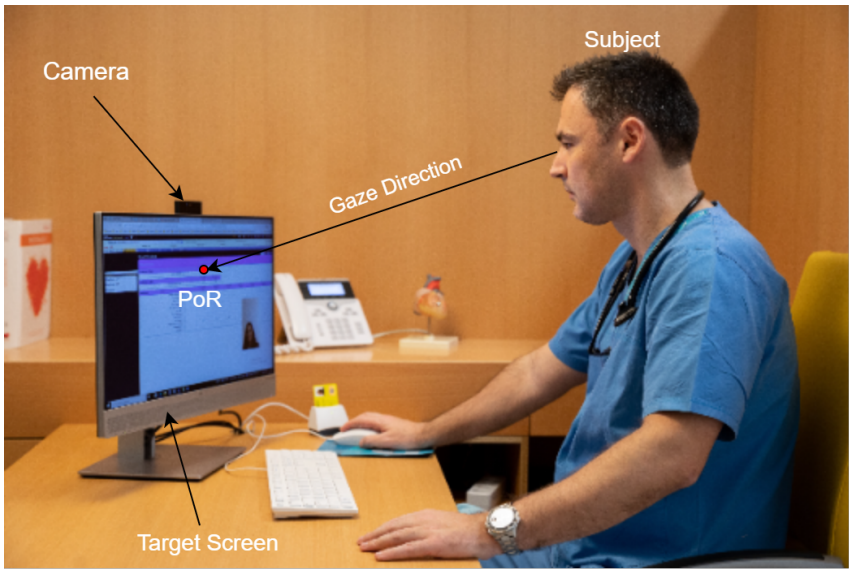


Figure 2.9: Typical Desktop Gaze Estimation Setup.

2D Gaze Estimation

In 2D gaze estimation, the task is to regress from the input image of the subject to a 2-dimensional on-screen gaze location $\boldsymbol{p}$ as $\boldsymbol{p} = f(\boldsymbol{I})$, where f is the regression function and $\boldsymbol{I}$ is the input image [29, 42, 56, 55, 33]. The point $\boldsymbol{p}$ is typically defined in the screen coordinate system and the screen is defined as a virtual plane in the camera coordinate system. The regression function typically needs the input image and additional information like 3D head pose [28] and face bounding boxes locations [29].

One limitation of 2D gaze estimation comes from the fact that it assumes that the screen and camera have fixed positions in relation to each other. In other words, the target screen plane is fixed in the camera coordinate system. Consequently, 2D gaze estimation tasks do not allow for free camera movement after training the system. Another limitation for 2D gaze estimation comes from another assumption that the camera's intrinsic parameters are always the same. Therefore, a trained regression function cannot be applied to different cameras.

3D Gaze Estimation

In 3D gaze estimation, the task is to regress from the input image $\boldsymbol{I}$ to a 3-dimensional vector $\boldsymbol{g}$, as $\boldsymbol{g} = f(\boldsymbol{I})$, in the camera coordinate system. The function f, optionally, takes the 3D head pose as an additional input apart from the input image [54]. The gaze direction output is a vector typically represented as a unit vector originating from a point of reference in the image, like center of the eyes [42, 46, 33, 11] or center of the face [28, 55, 56].

Data Pre-processing

Data pre-processing in appearance-based gaze estimation consists mainly in two tasks: face and eye detection and data rectification. Face/eye detection is needed to prune all unnecessary information from raw images since estimating gaze from raw images will lead to the use of more computational resources

and increase the influence of environmental factors which are not important for the gaze estimation task. Therefore, face alignment methods are applied to the raw images to obtain the facial landmarks and then crop the face/eye region of the subject. A list of some of the modern face alignment methods used today is presented in Table 2.1.

Table 2.1: Face Alignment Methods

Ref.	Method	Year
[27]	Dlib	2014
[53]	MTCNN	2016
[3]	OpenFace	2018
[16]	3DDFA_V2	2020

Data rectification in appearance-based methods is needed to eliminate as much environmental variability that comes from unconstrained environments like head pose rotation, illumination and background changes. All these factors lead to an increase in complexity of eye appearance and consequently in the complexity of the gaze estimation task. The current data rectification methods concentrate in dealing with head pose and illumination in unconstrained environments.

Head pose can be decomposed into rotation and translation of the head in the camera coordinate system, it degrades eye appearance. In [46], a rectification method for eye and face images is proposed which performs a perspective transformation from a virtual camera image such that eye appearance variations caused by head pose changes are mostly cancelled out. This method eliminates the ambiguity of different head poses, in Figure 2.10 we illustrate the visual effect the perspective transformation has on the images. On the other hand, Illumination changes are dealt with by performing histogram equalization on grey-scale images, therefore methods usually take grey-scale instead of RGB images.

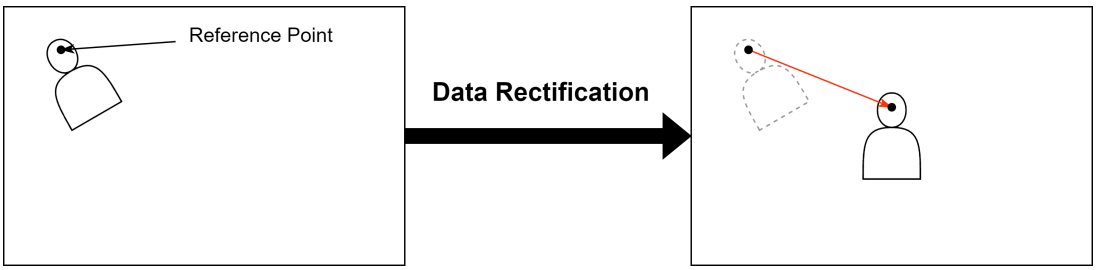


Figure 2.10: Illustration of the visual effect of the data rectification method from [46] on the images. The reference point refers to the center of the face, obtained with one of the face alignment methods from Table 2.1

Data Post-processing

Post-processing of gaze estimation results is often needed depending on the application using the gaze estimation method. Some applications might need 2D gaze estimates if they are evaluating screen-

based interactions, others might only need 3D gaze estimates. Consequently, the conversion between 3D gaze directions and 2D estimates/*PoG*s is a very common post-processing step in gaze estimation applications. To perform the conversion between 3D and 2D gaze estimates the extrinsic parameters of the camera explained in Section 2.2 are needed to convert the estimates from Camera Coordinate System to Screen Coordinate System and vice-versa. Here we introduce the common rountine used to convert between 3D gaze and 2D gaze, the notation and symbols used are summarized in Table 2.2 and Figure 2.11 illustrates them.

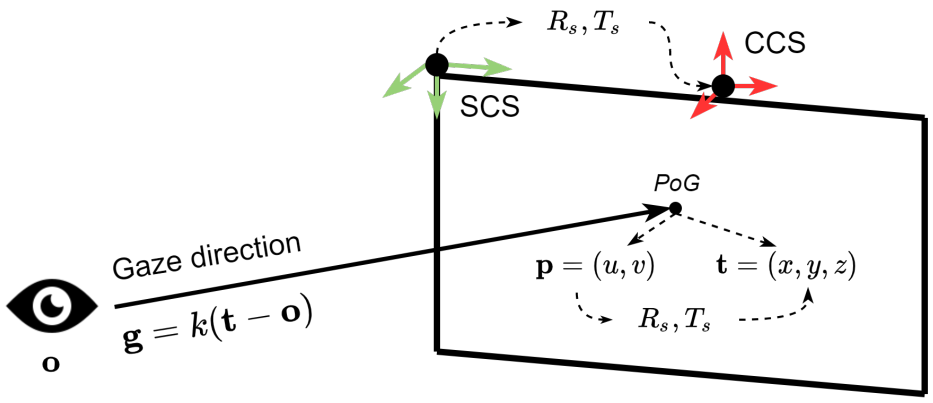


Figure 2.11: Illustration of a typical desktop gaze estimation task with 3D space representation of the symbols in Table 2.2

Symbol	Meaning
CCS	Camera Coordinate System
SCS	Screen Coordinate System
$\rho \in \mathbb{R}^2$	$\rho = (u, v)$, 2D location of gaze targets on SCS
$t \in \mathbb{R}^3$	$t = (x_t, y_t, z_t)$ 3D location of gaze targets in the CCS
$g \in \mathbb{R}^3$	$g = (g_x, g_y, g_z)$ gaze directions in the CCC
$o \in \mathbb{R}^3$	$o = (x_o, y_o, z_o)$ origin of gaze direction in the CCS
$R_s \in \mathbb{R}^3$	Rotation Matrix of SCS w.r.t to CCS
$T_s \in \mathbb{R}^3$	Translation vector between CCS and SCS
$n \in \mathbb{R}^3$	$n = (n_x, n_y, n_z)$ Normal vector of the 2D Screen plane

Table 2.2: Gaze conversion Symbols

The goal of converting from 2D gaze to 3D gaze is to take the 2D gaze estimate, $\rho = (u, v)$, and obtain the 3D gaze direction $g = (g_x, g_y, g_z)$. First, from the 2D gaze estimate $\rho = (u, v)$, we obtain the 3D location of ρ in the CCS denominated of $t = (x_t, y_t, z_t)$. The 3D location t can be computed as $t = R_s[u, v, 0]^T + T_s$, where the additional 0 is the z coordinate of ρ in the SCS. The gaze origin o is obtained from the face alignment method used in the pre-processing phase. With t and o, the gaze

direction vector g is computed as the vector between these two points given by equation (2.12):

$$g = \frac{t - o}{\|t - o\|} \tag{2.12}$$

On the other hand, converting from 3D gaze to 2D gaze is to obtain 2D gaze estimate ρ given 3D gaze direction g. The gaze origin o is obtained from the face alignment method. First, the intersection between gaze direction g and the screen plane in the CCS is computed to give the 3D location t. Then t is converted to p using the cameras extrinsic parameters. The screen plane equation can be obtained from the extrinsic parameters R_s and T_s where R_s gives the normal vector of the screen plane and T_s gives a point in the screen plane:

$$R_s = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \end{bmatrix} \tag{2.13}$$

$$n = (r_{12}, r_{22}, r_{32}) = (n_x, n_y, n_z) \tag{2.14}$$

$$T_s = [t_x, t_y, t_z]^T \tag{2.15}$$

and the plane equation is given as:

$$n_x x + n_y y + n_z z = n_x t_x + n_y t_y + n_z t_z \tag{2.16}$$

Given a gaze direction $g = (g_x, g_y, g_z)$ and gaze origin $o = (x_o, y_o, z_o)$, the gaze vector in the CCS is written as:

$$\frac{x - x_o}{g_x} + \frac{y - y_o}{g_y} = \frac{z - z_o}{g_z} \tag{2.17}$$

The intersection t is obtained by solving (2.16) and (2.17). The 2D coordinate ρ is obtained from $(u, v, z) = R_s^{-1}(t - T_s)$, where z is usually 0 and $\rho = (u, v)$ is the *PoG* on screen.

2.3.3 Deep Learning based Appearance-based Gaze Estimation

Initial appearance-based methods usually learn subject specific mapping functions requiring time consuming personal calibration routines. Appearance-based gaze estimation faces many challenges, head motion and subject differences in unconstrained environments are among the biggest ones. Performance of conventional appearance-based methods drops significantly in unconstrained environments where head pose is not fixed and subject is not limited to one specific person. Several approaches were proposed to try and improve performance against these problems [46, 24, 34]. However due to weak fitting ability, conventional appearance-based methods were not able to handle these challenges.

Deep learning techniques like convolutional neural networks (CNNs), explained in Section 2.1.2 are used in a variety of computer vision tasks with great success, naturally, they were applied in appearance-based gaze estimation as well. Zhang *et al.* proposed the first CNN-based gaze estimation method [54]. They used a simple CNN to regress gaze direction directly from eye images and its performance

surpassed the majority of conventional appearance-based methods. Subsequently, several extensions and improvements on CNN-based gaze estimation methods were published. Currently, the field of appearance-based gaze estimation with deep learning techniques is research hotspot with several works being published each year, Yihua Cheng *et al.* published and in depth review and benchmark of deep learning powered appearance-based gaze estimation [9]. Two major state-of-the-art gaze estimation methods proposed are the *MPIIFaceGaze* [55] and Gaze360 [28].

Gaze Estimation Datasets

Along with the several different deep learning architectures creted to perform the gaze estimation task [29, 33, 28, 42, 56, 54, 55, 11], several gaze estimation datasets to train those architectures were also proposed. The more widely used datasets today are represented in Table 2.3, many of these are geared with desktop or smartphone gaze tracking in mind. They are captured with a static camera setup [12, 45, 42, 54, 55] or with a camera integrated into a mobile device [29]. The static approach leads to higher accuracy but is more limited in terms of variation of illumination and motion blur. The datasets geared towards mobile gaze tracking offer more variety in number and variety of subjects. Some datasets also contain additional information like the 3D head pose [54, 45, 12, 46].

Table 2.3: Datasets

Ref.	Dataset	Type of content	# Subjects	#Samples	Gaze
[54]	MPIIGaze	Face + Eye crops	15	213,659	3D
[28]	Gaze360	Face + Full body	238	172,000	3D
[12]	EYEDIAP	Face + Eye crops + Depth data	16	62,500	3D
[46]	UT Multiview	Face +Eye crops	50	64,000	3D
[45]	Columbia	Face	56	5,800	3D
[11]	RT-GENE	Face + Wearable device img + Depth data	15	122,533	3D
[42]	EVE	Face + Screen Content	54	12,308,334	3D
[29]	GazeCapture	Face	1,450	2,129,980	2D

The Columbia dataset [45] consists of 5880 images from 56 participants. It covers 5 different head poses and 21 gaze directions. However it has a limited variation in appearances. It was created with the idea of *gaze locking* in mind, which consists of sensing eye contact directly from an image in a passive, appearance-based manner. It also did not need any active illumination (like infra-red illumination), which was something commonly used at the time and reduced accessibility of gaze tracking methods.

The EYEDIAP dataset [12] is composed of data obtained from 16 people in a total of 94 sessions with a Microsoft Kinect and HD camera synchronized with 5 leds. Participants sat in front of the setup and looked at continuous and discrete targets on a computer screen. The EYEDIAP dataset was proposed

as the first benchmark dataset for gaze estimation from remote RGB and RGB-D images. It allowed for different methods to be compared and identify the advantages and disadvantages of each method.

The UT Multiview dataset [46] is composed of video sequences of 50 participants from 8 different views. It was created with the intent of eliminating the requirement for person- and session-dependent training for appearance-based 3D gaze estimation method. It has a wide variety of head poses, gaze directions and subjects.

The *MPIIGaze* dataset [54] was created for ”in the wild” gaze estimation when most of the existing datasets, at the time, contained low variety in head poses variation and illumination conditions. The EYEDIAP and the UT Multiview datasets were the only ones that had significant head pose variety, yet they lacked variety in illumination conditions since they were captured under laboratory conditions. The *MPIIGaze* dataset contains 213,659 images from 15 participants collected during natural everyday laptop use. It is significantly variable in relation to appearance and illumination.

The GazeCapture [29] is a large scale dataset made especially for *PoG* estimation in mobile devices, but it can be extended to desktop environments. There was a need for a dataset with a large number of subjects, with high head pose and illumination variety, since the existing datasets did not provide enough subject variety. Consequently, crowdsourcing was used to build the dataset, addressing the lack of variety problem. The GazeCapture dataset is composed of 1450 people with unconstrained head motion and a wide range of backgrounds and illumination.

The RT-Gene dataset [11] is geared for dealing with high subject to camera distances and high variation in head poses and eye angles. It was the first work to try and address the problem of high subject to camera distances. To build the dataset, RGB-D data of the subject wearing a mobile eye tracker is recorded. To avoid the changes in human appearance caused by the mobile eye tracker, semantic *inpainting* was used in the regions covered by the eye tracking glasses.

The Gaze360 dataset [28] is a large-scale dataset for robust 3D gaze estimation in unconstrained environments. Most of the datasets developed until then were geared towards desktop or smartphone gaze tracking, all of them had a static setup with a fixed recording setup. For this dataset, the ground-truth was calculated by placing a panoramic camera at the center of the scene between the subjects and a large rigid target marked with an AprilTag [40]. The Gaze360 dataset focused on containing a wide range of indoor and outdoor environments. Since it was not constrained to a static setup to capture data, the Gaze360 dataset has a variation of head pose much higher than any of the previous datasets.

The EVE dataset [42] is collected from 54 participants from 4 camera views. It contains over 12 million frames adding up to 105 hours of video data. A unique particularity of the EVE dataset is that along with the subject video data, the viewer’s content on the screen was captured simultaneously and denominated as *visual stimuli*. The intent is to refine the initial estimate by using the *visual stimuli* presented to the subject at the estimate’s time. Currently, the EVE dataset is the largest and most recent gaze estimation dataset.

Full-Face Gaze Estimation with a Spatial Weights CNN Architecture - *MPIIFaceGaze*

The authors of the work in [55] wanted to evaluate how valuable the information contained in the facial appearance of the subject, apart from the eyes, would be for a gaze estimation model. Therefore they propose a spatial weights CNN architecture for full-face appearance-based 3D and 2D gaze estimation, denominated *MPIIFaceGaze*. The input of the CNN would be a full face image, and the output would be a 3D gaze direction. The representation of the architecture can be seen in Figure 2.12. The objective

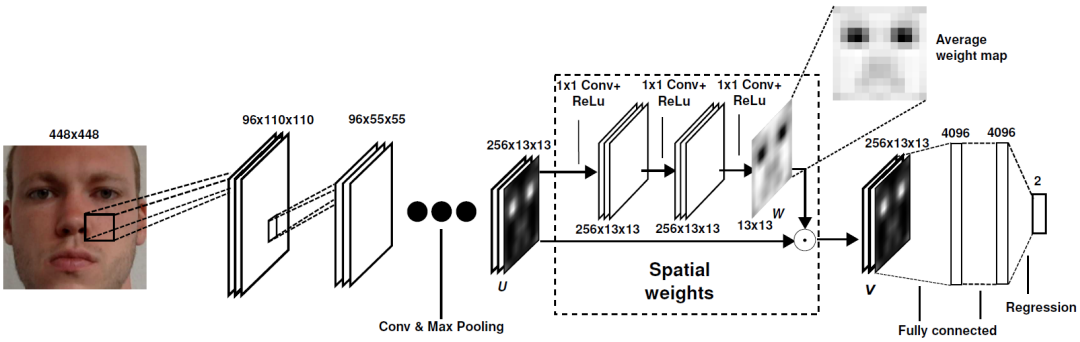


Figure 2.12: Spatial Weights CNN for full-face appearance-based estimation from [55].

with the spatial weights mechanism is to force the network to learn and understand that different regions of the face have different importance for the estimation task in a given sample. The mechanism consists of an additional 3 Convolutional layers with 1x1 kernel size followed by ReLU activation layers. The input of the additional convolutional layers is the activation tensor $\boldsymbol{U}$ with size $N \times H \times W$ $(256 \times 13 \times 13)$. N is the number of feature channels, H and W are the weight and width of the output. The activation tensor is passed through the additional layers to yield the spatial weights matrix $\boldsymbol{W}$. After this, a element-wise multiplication described by (2.18) between $\boldsymbol{W}$ and $\boldsymbol{U}_c$, where $\boldsymbol{U}_c$ is the c-th feature channel of $\boldsymbol{U}$, is performed yielding $\boldsymbol{V}$.

$$\boldsymbol{V}_c = \boldsymbol{W} \odot \boldsymbol{U}_c \tag{2.18}$$

The weighted activation tensor $\boldsymbol{V}$ is then fed into the fully connected layers of the network leading to the final estimate of the gaze direction. This method is implemented by the *OpenGaze* software toolkit proposed in [56]. *OpenGaze* provides a full gaze estimation pipeline for 3D gaze estimation. Additionally, when provided with the intrinsic and extrinsic camera parameters, *OpenGaze* can project the 3D gaze directions into 2D estimates on the target screen. *OpenGaze* is the first software toolkit developed for appearance-based gaze estimation and interaction.

Gaze360

The work in [28] aimed to provide robust gaze estimation in unconstrained environments with a wide variety of head poses and illumination conditions. To achieve this, they propose both a dataset and a deep learning gaze estimation model, both denominated of Gaze360. The Gaze360 model is a video-based gaze tracking model using an RNN architecture, more specifically, bidirectional Long-Short Term Memory (LSTM) [14]. The objective behind the use of a RNN architecture is that gaze is a continuous

signal and it is dependent on the past and future and modelling that dependency is needed. To model the temporal dependencies, the bidirectional Long-Short Term Memory architecture from [14] was chosen. The bidirectional LSTM architecture has two layers of LSTM modules, one to model dependencies going forwards in time and another to model dependencies going backwards in time.

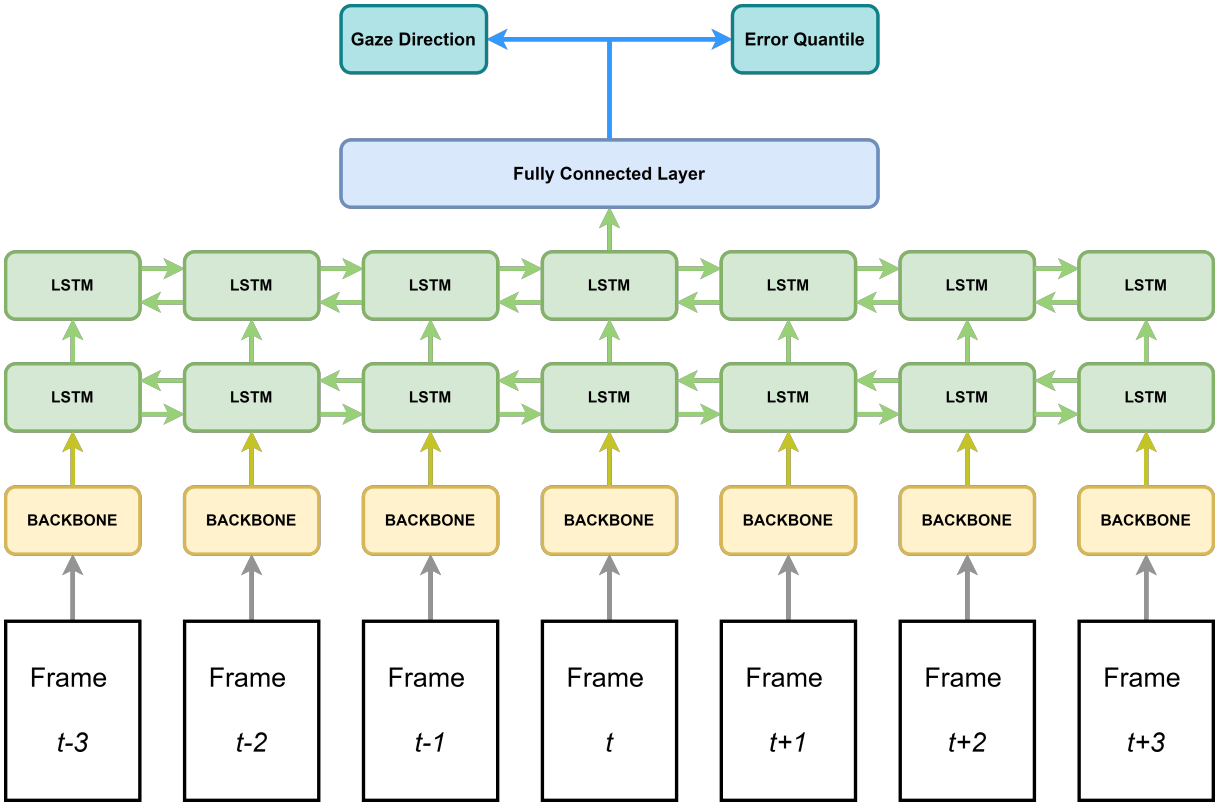


Figure 2.13: **Gaze360 model**: The model obtains 7 consecutive frames as input: the target frame *t*, the 3 previous frames and the 3 next frames. Each frame is passed through the backbone network of the model, the bidirectional LSTM layers and finally the Fully connected layer to give the gaze direction and quantile error results

In Figure 2.13, the full Gaze360 architecture is illustrated. The model utilizes sequences of 7 frames with the target frame being the central frame of the sequence. In essence, to obtain the gaze estimate for one specific frame the model looks at the frame in question as well as the previous 3 frames and next 3 frames. Before being fed into the backbone network, each frame is pre-processed which consists of cropping the image according to the bounding boxes of the face obtained with the a face alignment method (examples in Table 2.1) and performing an histogram equalization to deal with illumination variations. The pre-processed images are then fed to the backbone networks, a ResNet-18 CNN [21] pretrained on the ImageNet dataset. The backbone network converts the facial images into high-level features with dimensionality of 256. The features are then sent into the 2 layers of bidirectional LSTMs, one layer processes the sequence of features in one direction and the other on the reverse direction. After being processed, all the feature vectors are concatenated and fed to the final fully connected layer which regresses them into the 3D gaze direction and an error quantile estimation. The error quantile estimation gives us the error bounds of the gaze estimate, it is a way to evaluate the confidence the model has that its estimate is accurate and it is very useful when estimating gaze in unconstrained

environments due to all the different factors that hinder the estimation accuracy.

2.4 Related Work

As mentioned in 1 the majority of works that analyzed doctor's gaze behaviour have collected gaze data manually in a cumbersome and slow process. However, in recent years, some works proposed implementations of automated gaze estimation approaches, in these section we will explain them.

The work in [48] proposed a gaze classification approach for doctor's gaze using CNNs. The CNN architecture used the VGG-16 architecture from [44] pre-trained on the ImageNet dataset as backbone and added to the backbone 1 Global Max Pooling layer, 1 Dropout layer and 5 fully connected layers which were then fine tuned with a dataset of raw videos from clinical interactions developed by the authors of the study. This dataset consisted of a set of 101 clinical interactions involving 10 doctors and 101 patients. Each interaction was comprised of 3 videos captured at the same time from 3 different cameras, a *Patient-Centered* camera focused on the patient, a *Doctor-Centered* camera focused on the doctor and a *Wide-frame* camera providing wide-view image of both the doctor and patient. The videos were annotated using the Noldus Observer XT Software [59]. In the end, the model proved to be very accurate with a 98.31% accuracy on the validation set and over 80% accuracy on the majority of independent hold out sets with unseen doctors and interactions. However, the need to use 3 different cameras with specific views of the consultation office makes this system unfeasible in many situations.

The work in [26] uses eye-tracking glasses to track the gaze of the doctor during face-to-face consultations. A total of 16 doctors seeing a total of 100 patients, each doctor seeing between 2 and 14 patients with the median being 6 patients. Doctor's gaze was measured with 3 different metrics: *face gaze duration*, *face gaze frequency* and *face gaze dwell time*. *Face gaze duration* corresponds to the total amount of time per minute the doctor spent looking at the patient, *face gaze frequency* corresponds to the amount of times per minute the doctor's gaze switched to the patient and *face gaze dwell time* is the time the doctor's gaze dwelled on the patients in each instance the doctor looked at them. The study concluded that there was a significant positive correlation between *face gaze dwell time* and *face gaze duration*, a significant negative correlation between *face gaze dwell time* and *face gaze frequency* and no correlation between *face gaze frequency* and *face gaze duration*. Additionally the study also concluded that the amount of face gaze present in the beginning part of the consultation had positive and significant association to the amount of face gaze present in the beginning of the consultation. It also found that the amount of time the doctor spends looking at the patient during a consultation decreases in the final parts of the consultation compared to the beginning of the consultation.

2.5 Conclusion

In this chapter, we went over the history of approaches used for gaze estimation focusing on deep learning appearance-based approaches. These deep learning powered approaches do not require special hardware (normally expensive) and are much more robust in unconstrained environments when com-

pared to other types of approaches. This flexibility combined with the state-of-the-art performance is what makes them, currently, a research hotspot in the area of gaze estimation. In addition, two different deep learning appearance-based methods have been described: the *MPIIFaceGaze* method and the *Gaze360* method. Both methods have state-of-the-art performance when evaluated against current gaze datasets [9]. However, performance in unconstrained environments could not be evaluated just by benchmark accuracies alone. The evaluation of performance in unconstrained environments is shown in the next chapter in Section 3.3.1.

In addition to the gaze estimation methods this chapter also explained the typical processing pipeline used for gaze estimation methods. This pipeline consists of data pre-processing and post-processing steps. Data pre-processing consists of tasks like face/eye detection, data rectification and histogram equalization. Data post-processing mainly consists of the conversion between 2D and 3D gaze estimates. For this study we will use *3DDFA_V2* [16] as our facial detection and landmarking method and to find the camera's extrinsic parameters we will use the mirror calibration method from [47].

Chapter 3

Experimental Design and Protocol

In this chapter, a description of the approach used is presented based on the work outlined in Chapter 2. Section 3.1 states the study's goal and presents the hypotheses being tested. Section 3.2 describes the data acquisition protocol. First, it describes the setups used in the consultation offices and then explains the interfaces the doctors were asked to interact with to record the consultations. Section 3.3 explains the gaze estimation pipeline used and how it is implemented, and additionally, it describes the comparisons used to select between *Gaze360* and *MPIIFaceGaze*. Section 3.4 explains the methodology used for the study.

3.1 Experimental Design

The goal of this project is to quantify and compare the amount of time a doctor is looking at the patient between virtual and face-to-face consultations. This will allow the verification of the preliminary observations made by the Luz Saúde group mentioned in Chapter 1. Thus the hypotheses we wish to test with our study are the following:

- **Hypotheses 0 (H0)**: The doctor spends the same time looking at the patient during face-to-face consultations and virtual consultations, this is the null hypotheses
- **Hypotheses 1 (H1)**: The doctor spends different amounts of time looking at the patient during face-to-face and virtual consultations

Therefore we implemented a gaze estimation system in the consultation environment to record and evaluate a doctor's gaze during the consultation and classify it according to whether the doctor is looking at the patient. For each doctor involved in the study, we recorded and evaluated a set of face-to-face and virtual consultations to obtain the percentage of time in the consultation where the doctor looked at the patient. The two sets of percentages would be compared against each other to understand if there were any significant statistical differences or tendencies between them. This would finally allow us to conclude about the influence of the virtual consultation environment on the gaze behaviours of the doctor during a consultation.

3.1.1 Source Population and Sample Size

The source population for this study consists of 14 doctors divided among 4 different medical specialties. For each doctor, 20 face-to-face and 20 virtual consultations were recorded and analyzed. Choosing doctors among 4 medical specialties increases the scope of our study, at the expense of some robustness in the data. However, it was deemed more important understanding the impact of the virtual consultation environment on multiple types of medical specialties instead of focusing in just one. The 4 medical specialties chosen were:

- **Family and General Medicine** - 4 doctors
- **Endocrinology** - 3 doctors
- **Gynecology**/**Obstetrics** - 4 doctors
- **Neurology** - 3 doctors

Every doctor gave his/her informed consent, shown in Appendix A to record their face during consultations (no audio recorded), where no patient information was recorded. Patients were informed of the project in a verbal manner and consented participation.

3.2 Data Acquisition Procedure

The acquisition of data for this study consisted in the recording of 20 face-to-face consultations and 20 virtual consultations for each participating doctor. The use of the hospital's camera and systems was not allowed due to a variety of privacy and security reasons. Hence an extra camera and computer (running a GUI for the doctor to interact with) were setup in the consultation offices. In the following sections we will describe the 2 different consultation offices, the limitations that came with them and the graphical interface developed so that the doctors were able to record the consultations without the need of a third party being present at the consultation. For privacy purposes, only the video feed from the camera pointing to the doctor was saved. No audio, whether from the patient or the doctor, was recorded.

3.2.1 Consultation Environments

The study considers two consultation environments: the conventional consultation room and the virtual consultation room. The conventional consultation room allows the doctor to perform face-to-face consultations, see Figure 3.1a. During an face-to-face consultation, the patient is in the room with the doctor and the patient's health records are shown on the screen. The virtual consultation room is specially designed for virtual consultations, see Figure 3.1b. It has two different screens, one where the patient's video feed is shown and another which shows the health records of the patient.

In each of the consultation rooms, a webcam was set up in order to look directly at the doctor's face, this provides the input to the gaze estimation model to obtain the gaze direction/*PoG* of the doctor. In the conventional consultation room the camera was placed on top of the only computer screen and in the

(a) Conventional consultation room (b) Virtual consultation room

Figure 3.1: Types of consultation rooms: Each room is was set up with an extra camera and computer to record the doctor's image during the consultation.

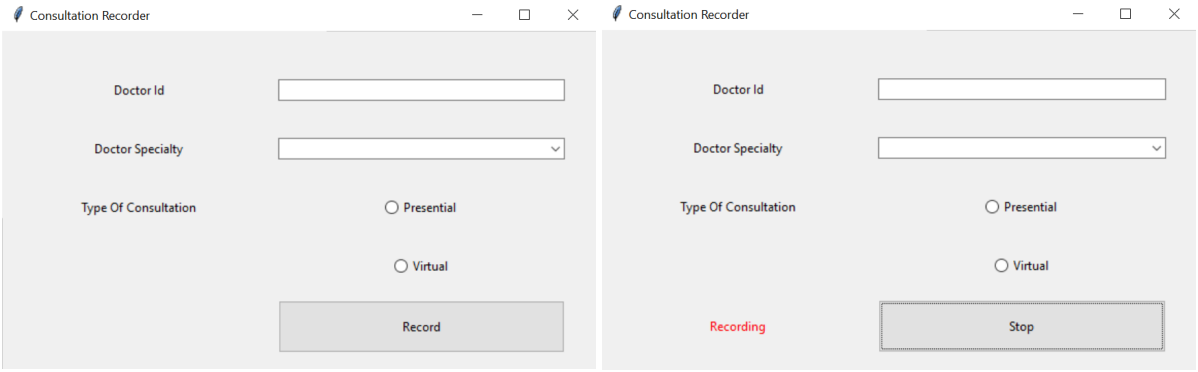
virtual consultation rooms the camera was placed on top of the screen which shows the patient health records.

For each of the consultation rooms, the patient position was recorded as being on the left of the screen or on the right of the screen, which depended on the consultation room layout, this was done to enable the classification of the 2D gaze estimates later when processing the data. Due to the pandemic doctors were obligated to wear a mask during face-to-face consultations, thus the doctors were asked to use a special mask, shown in Figure 3.2. This mask does not cover the nose and mouth area of the face, which improves face detection and gaze estimation accuracy, details on the evaluation of the special and conventional types of masks are presented in Section 3.3.1.

Figure 3.2: Special mask worn by the doctor's during face-to-face consultations

3.2.2 Recording Interface

A simple GUI, illustrated in Figure 3.3, was developed for the doctors to interact with, and record the consultations. The development of the GUI was done using the *Tkinter* Python interface library [35]. The doctors' interaction with the GUI was resumed by clicking a button at the beginning of the consultation and end. At the beginning of the consultation, the GUI is in the state seen in Figure 3.3a. When the



(a) Before recording starts

(b) While recording

Figure 3.3: Recording GUI: The GUI had 2 different states to indicate to the doctors if it was recording or not. The 3 blank fields (*Doctor Id, Doctor Specialty, Type Of Consultation*) would be filled according to the doctor and environment before hand.

consultation starts, the doctor clicks the *Record* button to indicate the start of the consultation, which initiates the recording. After clicking the button, an *FFmpeg* command [49] is executed to start the recording the video from the webcam. Videos were filmed at a 720p resolution and 15 fps. Furthermore, the GUI changes from Figure 3.3a to Figure 3.3b, to indicate to the doctor that the recording is in progress. At the end of the consultation, the doctor clicks the *Stop* button to indicate the end of the consultation and terminate the *FFmpeg* execution. In the end, a *.mp4* video file (with no audio) with the consultation recording is saved for further analysis by the gaze estimation pipeline.

3.2.3 Extrinsic Camera Calibration Procedure

As mentioned in Section 2.2, a mirror-based extrinsic calibration method from [47] was used to compute the extrinsic parameters of the camera. The extrinsic calibration procedure required to take 9 pictures of the screen in relation to which we wanted to calibrate the camera.Additionally, the screen had to have attached a checkerboard pattern to serve as reference points for the calibration procedure, where the checkerboard squares' vertices would fulfill the role of reference points for the method. An example of the 9 pictures can be seen in Figure 3.4. The method takes as input the camera's intrinsic parameters, the 3D coordinates of the reference points in relation to the screen coordinate system, the 9 pictures of the checkerboard pattern from different angles and the pixel coordinates of each of the reference points in each picture. After taking the pictures, the checkerboard reference points pixel coordinates were obtained using the *OpenCV* library [8]. With the pixel coordinates obtained, we can calculate the rotation and translation matrices between the camera coordinate system and the screen coordinate system. After obtaining both matrices we can now convert 3D gaze estimates into 2D gaze estimates on the screen plane.

Figure 3.4: Example of the 9 pictures taken to serve as input of one of the extrinsic calibration procedures performed

3.3 Pipeline for autonomous gaze estimation

The objective of our gaze estimation task is to take a consultation video and analyze it frame by frame assigning each frame the doctor's *PoG* estimate. For this, a pipeline for autonomous gaze estimation was developed, composed of a face/landmark detection method, a gaze estimation model and a post-processing routine for converting the 3D gaze estimate to a *PoG* estimate. In addition to the pipeline, the mirror-based extrinsic calibration method from [47] was used to obtain the extrinsic parameters of the camera in relation to the screen needed to perform the gaze conversion.

The gaze estimation model for the pipeline was a choice between 2 deep learning appearance-based gaze estimation methods: *Gaze360* and *MPIIFaceGaze*. In the end, the *Gaze360* model proved to be the best for the task at hand, the process used to reach this conclusion is explained in 3.3.1. the pipeline structure and blocks are explained in 3.3.2

3.3.1 Performance in the Consultation Environment

The work in [9] performs a review and benchmark of deep learning appearance-based gaze estimation methods. From the methods evaluated, *MPIIFaceGaze* and *Gaze360* had state-of-the-art performance with *Gaze360* having a slight advantage over *MPIIFaceGaze*. However the implementation of a gaze estimation system in an unconstrained environment like the consultation environment comes with many challenges that are not covered by a comparison of the methods accuracy over current gaze estimation datasets.

The 3 biggest challenges identified in the consultation environment were the high variability of head poses, the use of glasses and the use of masks. Several practical tests were done with *MPIIFaceGaze* and *Gaze360* to evaluate these additional complexities when implementing such a system in an unconstrained environment. The test consisted of recording videos that would enable the comparison of the performance of these methods against these variables. The videos recorded consisted of the subject

simulating doctor behaviour during consultations. The robustness against head pose variability could be verified by checking how well each method could follow the subject's gaze. Some of the videos recorded had the subjects using glasses and masks to assess the impact of glasses and masks on performance.

Regarding robustness against different head poses and glasses, the *Gaze360* method far outperforms *MPIIFaceGaze*, Figures 3.5a and 3.5b clearly show how, when faced with non-frontal head poses, the *MPIIFaceGaze* method fails to follow the subject's gaze. When wearing glasses, the error is amplified. *Gaze360*'s ability to perform even with such extreme head poses is due to the *Gaze360* dataset [28], covering a wide range of possible subject head poses. The same can be said about subjects wearing glasses. The wide range of subject appearances in the *Gaze360* dataset helps the *Gaze360* method perform better in these unconstrained environments.

In terms of robustness against masks the *MPIIFaceGaze* method outperforms the *Gaze360* method. In Figure 3.5c, we can see how *Gaze360* cannot follow the eyes of the subject. This is because using a mask leads to the occlusion of most of the subject's facial appearance. Since there is not a large amount of labelled training data on mask-wearing subjects, both methods suffer a dip in performance. However, the *MPIIFaceGaze* spatial weights mechanism mitigates the mask effects by putting much more weight on the eye region of the face which is not covered by the mask. Therefore *MPIIFaceGaze* reacts much better to eye movements when the subject has a partially occluded face. During testing, it was found that if the subject used a mask that did not occlude the mouth and nose area of the face, the performance of the *Gaze360* would have improved performance, reducing the gap to *MPIIFaceGaze*.

(a) Robustness against Head Pose Variability

(b) Robustness against glasses

(c) Robustness against masks

(d) Robustness against glasses and mask

Figure 3.5: Screenshots of test recordings: The 3 vectors represent 3 different gaze directions: the output of *MPIIFaceGaze* (*red*), the output of *Gaze360* (*white*) and the average of both outputs (*green*)

The *Gaze360* method proved to be a more robust method against the unconstrained variables found in the consultation environment and capable of following the subject's gaze correctly in various situa-

tions. In Figure 3.5d, we have a screenshot of one of the final testing recordings with a doctor doing a mock consultation. In this scenario, the *Gaze360* method ended up performing significantly better. *MPIIFaceGaze* outperformed *Gaze360* when the subject was facing the screen/camera directly (the camera was placed on the top right corner of the screen). However, since the study's objective is to follow the doctor's gaze switching between the patient and the screen, the inability of *MPIIFaceGaze* to perform well against a wide range of head poses leads to the choice of *Gaze360* as the gaze estimation method for this study.

3.3.2 Pipeline

The pipeline constructed is composed of a pre-processing step (face/landmark detection), gaze estimation step and a post-processing step (conversion from 3D to 2D gaze estimates). For the face/landmark detection we use the 3DDFA_V2 from [16], which was chosen due to possessing state-of-the-art performance and keeping that accuracy when subjects were wearing a mask. A simple illustration is shown in Figure 3.6. The input of the method is the consultation recording while the output is frame-by-frame list of face bounding boxes (to locate the doctor's face in the image) and 2D landmark annotations (pixel coordinates) of the doctor's face. Since the method only outputs pixel coordinates there is no need for the camera's intrinsic parameters in this step.

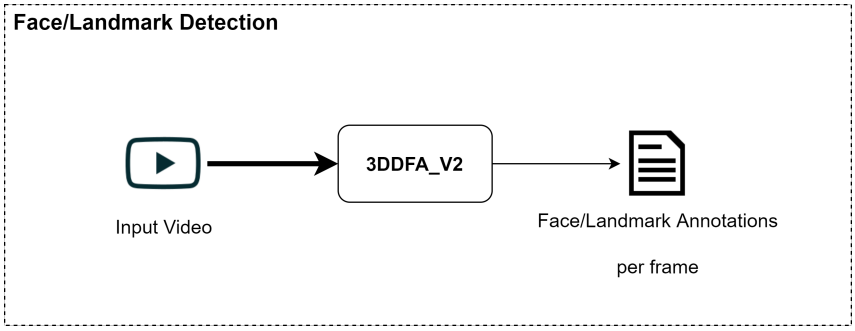


Figure 3.6: Face landmark detection input-output diagram

For the gaze estimation step, the Gaze360 model from [28] is used to extract the doctor's *PoG* during the consultation from the consultation video and the face/Landmark annotations provided by the previous step illustration of the gaze estimation section inputs and outputs is shown in Figure 3.7.

The post-processing consists in performing the conversion of the 3D gaze estimates to 2D gaze estimates according to the method explained in the *Data Post-Processing* component of Section 2.3.2. Its inputs and outputs are illustrated in Figure 3.8.

In the end, the pipeline performs the task of taking a consultation video and outputting the 2D gaze estimates of the doctor per frame. The full pipeline is illustrated in Figure 3.9.

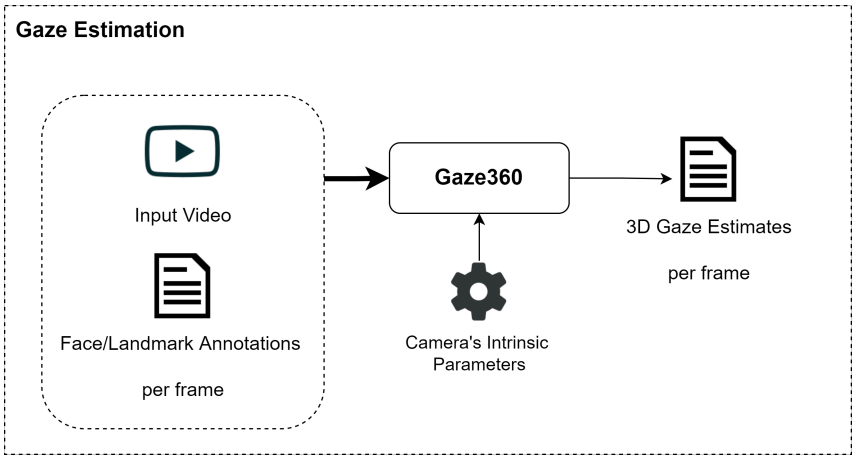


Figure 3.7: Gaze estimation input-output diagram

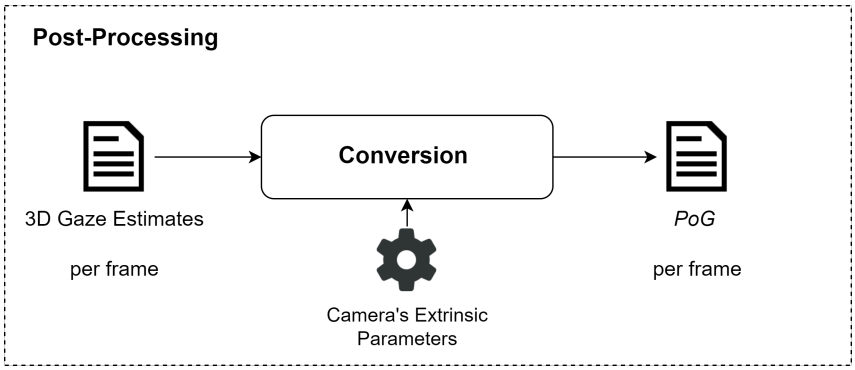


Figure 3.8: Post-processing input-output diagram

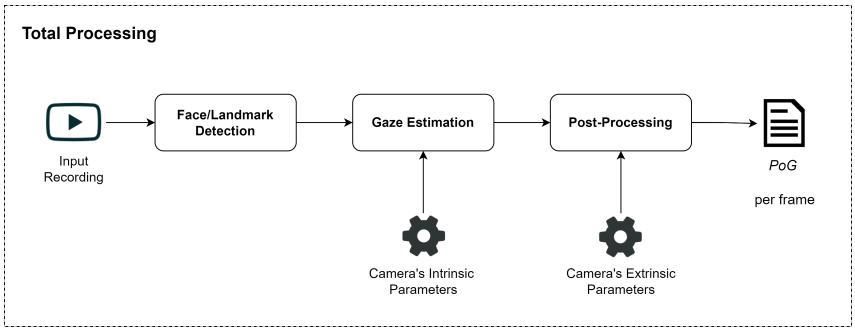


Figure 3.9: Pipeline used to process consultation videos into 2D gaze estimates

3.4 Data Classification and Analysis

The doctors' *PoG*s obtained from the gaze estimation pipeline are classified according to whether the doctor looks at the patient or the computer/keyboard. Then, with the classified *PoG*s, we can extract the percentage of time of the consultation during which the doctor looked at the patient denominated by *Patient%*. The *Patient%* statistic will be the target variable compared in our study between consultation environments. With the distributions of *Patient%* from both consultation environments, we can compare the doctors' gaze behaviour between them and assess if there are any statistically significant differences. Both the *PoG* classification and the statistical comparison processes are explained in this section.

3.4.1 *PoG* Classification

The classification consists of taking the frame by frame *PoG*s of each video and classifying them according to which zone of the screen plane they are located. The zones of the screen plane were defined according to the screen bounds and are illustrated in Figure 3.10. In total, 5 different zones were defined: *Above Screen*, *Keyboard*, *Screen*, *Right Of Screen* and *Left Of Screen*. Each *PoG* is assigned to one of these zones according to its location on the screen plane. Depending on the patient's position in the consultation environment, either the *Right of Screen* or the *Left of Screen* zones would be considered the *Patient* zone. In face-to-face consultations, when the doctor examines the patient in the back

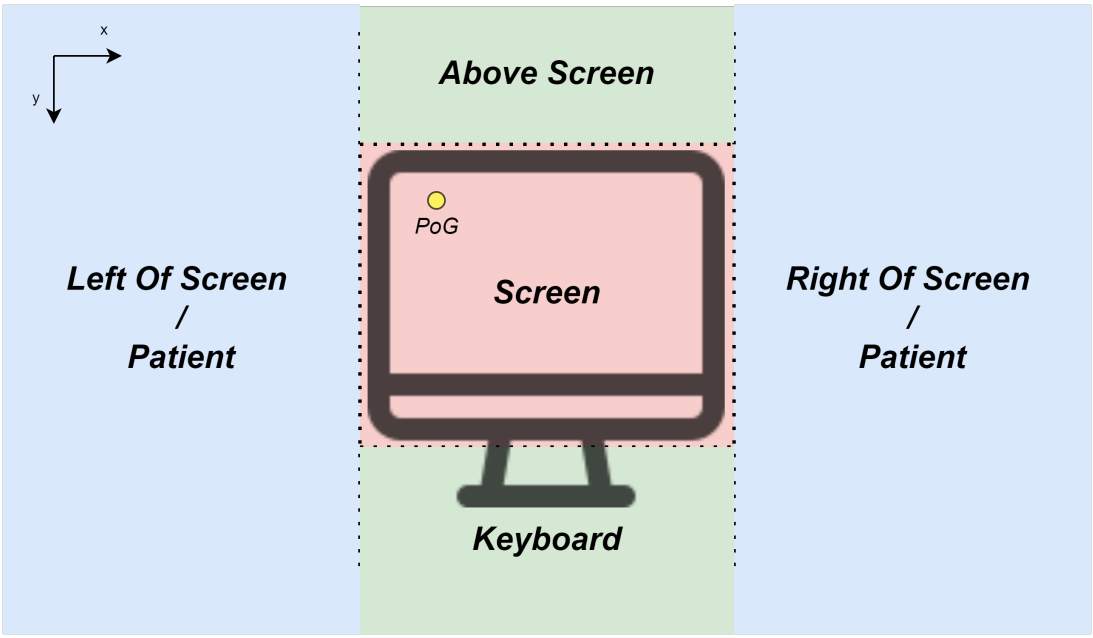


Figure 3.10: 5 Classification zones dividing the screen plane: The *Patient* zone would change depending on the consultation room and where the patient would be positioned in that specific consultation room

of the consultation room, the doctor's face goes out of the camera's view. This situation leads to the face detection step failing and no *PoG* being outputted for those frames. Therefore, to deal with these situations, we considered every frame where the doctor's face is not detected to be a *Patient* frame since the doctor would be interacting physically with the patient in the back of the consultation room. In the end, each of the frame-by-frame *PoG*s is assigned a classification. These classifications enable us to tell where the doctor is looking at any point in the consultation. Therefore, we can finally obtain the *Patient%* as the percentage of *PoG*s in the *Patient* zone.

3.4.2 Statistical Tests

To compare the percentage of time the doctor is looking at the patient, during both types of consultations, we use the percentage of time the doctor spent looking at the patient during the consultation, which we will denominate of *Patient%*, and check for statistically significant differences between virtual and face-to-face consultations.

First, the data distributions are always assumed to be non-normal distributions due to the low sample

size (20 data points for each distribution). Therefore the test chosen needs to be a non-parametric test. We are testing for differences between groups in different conditions (face-to-face or virtual consultations) and with different participants (different patients). Therefore, the adequate non-parametric test for this situation is the Mann-Whitney U test [36], a non-parametric test which does not have any prior assumptions over the data distribution. The Mann- Whitney U test will tell whether there are statistically significant differences between both data distributions through the p-value. If the p-value is smaller than 0.05, there is a statistically significant difference. In cases where there is a statistically significant difference, we will also look at the size effect using the Cohen's d size effect [10]. The effect size provides a measure of how far are the means from each group in relation to data variability and it is used to assess how big of a change did the consultation environment have on the *Patient%* statistic. If the d is below 0.2, then the size effect is considered small, if it is between 0.2 and 0.8 is considered medium and above 0.8 is considered large. With these tests we are able to test the hypotheses mentioned in 3.1 about the percentage of time the doctor spends looking at the patient.

We will perform a Mann-Whitney U test for each doctor between his virtual consultation and face-to-face consultation distributions to assess whether any specific doctor has different behaviours in both types of consultation environments. Additionally, for each medical specialty, we will group all the face-to-face and virtual results of doctors from that specialty and perform the Mann-Whitney U test on the joint data distributions, this will allow us to get a more general view on the differences in behaviour between types of consultations within the medical specialty. These tests will allow us to say with confidence whether any doctor had significant changes in gaze behaviour between consultation environments. However, in the case of the test by medical speciality, where we gather data from all the doctors, we want to see what kind of pattern emerges. This grouping provides a speciality analysis to see if there are specific patterns per speciality that require further data recording.

Chapter 4

Results and Discussion

In this chapter, we will present the results of this study. The chapter is divided into five sections, one section for each medical specialty and one section discussing the results. In total, 13 doctors participated in the study, they are presented as DX where X is the unique number given to the doctor to distinguish between them. Every doctor recorded 20 face-to-face and 20 virtual consultations, each one of these recordings was processed to obtain the *Patient%* metric, resulting in each doctor possessing two distinct distributions: the face-to-face distribution and the virtual distribution, which correspond to the *Patient%* distributions in each type of consultation environment. In the medical specialty sections, each doctor's results are presented individually. The individual doctor results are composed of three parts: the medians of the data distributions (Med_f and Med_v for the face-to-face and the virtual distributions respectively), the Mann-Whitney U test *p-value* plus size effect calculations (when a statistically significant difference is found) and the violin plot of the data distributions against each other. In this work, the violin plot shows, in addition to the distributions' shapes, all the data points of each doctor (represented with a black line). Additionally, the results of the joint data from the doctors of each specialty provide a more general overview of the medical specialty, are also shown.

4.1 Gynaecology/Obstetrics

In the Gynaecology/Obstetrics medical specialty, one doctor presents a statistically significant difference between *Patient%* distributions in face to face and virtual consultations, one doctor presents a tendency to have higher *Patient%* in virtual consultations and two doctors present no significant differences or tendencies between face-to-face and virtual consultations. The summary statistics of all doctors are shown in Table 4.1 and Figure 4.1 shows the violin plots of each doctor distributions.

Doctor D6's Mann-Whitney U test results in a *p-value* of 0.88, meaning no statistically significant difference exists between the face-to-face and the virtual distributions. Looking at Figure 4.1, we see that D6 face-to-face consultations' *Patient%* are more concentrated around the 50% mark, while the virtual consultations present a much wider range of values. This, combined with the face-to-face and virtual distribution medians (Med_f and Med_v) being almost identical, 51.8% and 51.9% respectively,

Table 4.1: Gynaecology/Obstetrics results and summary statistics

Doctor	Med_f	Med_v	*p-value*	Cohen's *d*
D6	51.8%	51.9%	0.88	-
D9	52.3%	58.9%	0.06	-
D12	55.7%	75.1%	$1.3e-5$	0.69
D14	47.9%	43.4%	0.56	-
Joint	51.3%	57.5%	$0.7e-2$	-

leads us to conclude that D6 presents almost no differences between both consultation environments.

Doctor D9's Mann-Whitney U test results in a *p-value* of 0.06, meaning no statistically significant difference exists between both distributions. However, looking at Figure 4.1, we see that D9's virtual distribution is slightly higher than the face-to-face distribution. Also confirmed by the distribution medians in Table 4.1, where Med_v is higher, 58.9%, than Med_f, 52.3%. These results show that D9 presents a tendency to look more at the patient during virtual consultations than in face-to-face consultations even though it is not statistically significant.

Doctor D12's Mann-Whitney U test results in a *p-value* of 0.000013, meaning that there is a statistically significant difference. In addition, the Cohen's *d* effect size is equal to 0.69, meaning that there is a medium effect size when changing consultation environments. This significant difference is further confirmed when looking at the violin plot of Figure 4.1 and at the Med_f and Med_v medians, 55.7% and 75.1% respectively. Therefore, we can say that D12 does look more at the patient in virtual consultations as opposed to face-to-face consultations.

Doctor D14's Mann-Whitney U test results in a *p-value* of 0.56, meaning no statistically significant difference exists between the face-to-face and the virtual distributions. D14's results are very similar to D6's in the sense that the violin plots show the same characteristics in both doctors. Additionally, the face-to-face median is slightly higher than the virtual median, 47.9% and 43.4% respectively. Therefore, when looking at the *p-value* and at Figure 4.1 we conclude that D14, just like D6, has almost no differences when changing between face-to-face and virtual consultations.

The Mann-Whitney U test on the joint data distribution of all doctors results in a *p-value* of 0.007, meaning there is a statistically significant difference between the face-to-face and the virtual distributions of the four doctors as a whole. Looking at Table 4.1, the median of the face-to-face distribution, 51.3%, is lower than the virtual distribution's median, 57.5%. However, since half the doctors present no differences between and D9 only presented a tendency, this leads us to believe that the majority of this difference in the joint test comes from the D12 doctor which had around a 20% difference between distribution medians. Therefore, to reach a conclusion about the effect of the virtual consultation environment in the Gynaecology/Obstetrics as a whole we would need more data and participating doctors. Nonetheless, these are promising initial results.

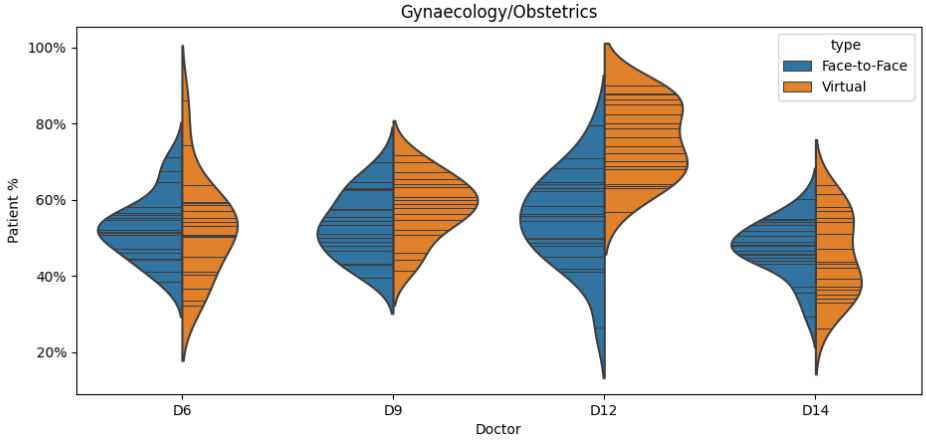


Figure 4.1: Gynaecology/Obstetrics doctors violin plots

4.2 Endocrinology

In the Endocrinology medical specialty, two doctors present a statistically significant difference between *Patient%* distributions in face to face and virtual consultations and one doctor presents a tendency to have higher *Patient%* in virtual consultations. The summary statistics of all doctors are shown in Table 4.2 and Figure 4.2 shows the violin plots of each doctor distributions.

Table 4.2: Endocrinology results and summary statistics

Doctor	Med_f	Med_v	*p-value*	Cohen's *d*
D3	37.7%	40.8%	0.11	-
D10	27.7%	42.2%	0.01	0.40
D15	44.6%	58.2%	$1.6e-5$	0.68
Joint	37.3%'s	50.9%	$0.7e-5$	-

Doctor D3's Mann-Whitney U test results in a *p-value* of 0.11, meaning no statistically significant difference exists between both distributions. However, we can note a tendency for higher *Patient%* in virtual consultations. When looking at Figure 4.2, we see that D3's virtual distribution values are slightly higher than the face-to-face distribution, further confirmed by the distribution medians in Table 4.2, where virtual consultations have a higher median, 40.8%, than the face-to-face distribution, 37.7%. Therefore, taking all the results into account, we can say that D3 presents a tendency to look more at the patient during virtual consultations than in face-to-face consultations.

Doctor D10's Mann-Whitney U test results in a *p-value* of 0.01, meaning that there is a statistically significant difference between the face-to-face and virtual distributions. Additionally, the Cohen's *d* effect size of 0.40 indicates a medium sized effect when changing consultation environments. The violin plot of D10, shown in Figure 4.2 shows the clear difference in the *Patient%* metric between distributions, with the virtual distribution having its peak at a higher value than the face-to-face distribution. This significant difference is further confirmed by the medians in Table 4.2, with the virtual distribution median being

42.2%, much higher than the face-to-face distribution median of 27.7%. Thus, we can safely say that D10 looks more at the patient during virtual consultations than in face-to-face consultations.

Doctor D15's Mann-Whitney U test results in a *p-value* of 0.000016, meaning that there is a statistically significant difference. In addition, the Cohen's *d* effect size is equal to 0.68, meaning that there is a medium effect size when changing consultation environments. This significant difference is further confirmed when looking at the violin plot of Figure 4.2 and at the Med_f and Med_v medians, 44.6% and 58.2% respectively, amounting to a difference of around 14%. Therefore, we can say that D15 looks significantly more at the patient in virtual consultations as opposed to face-to-face consultations.

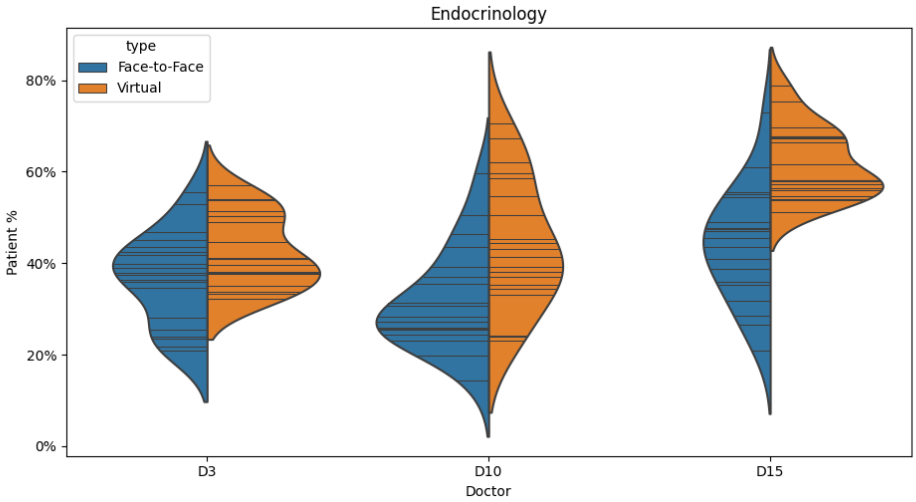


Figure 4.2: Endocrinology doctors violin plots

The Mann-Whitney U test on the joint data distribution results in a *p-value* of 0.000007, meaning there is a statistically significant difference between the distributions of the three doctors as a whole. Looking at Table 4.2, we can further confirm that the median of the face-to-face distribution is indeed lower than the virtual distribution median, 37.3% and 50.9% respectively. These initial results are very promising and indicate that the Endocrinology medical specialty has a tendency to look more at the patient during virtual consultations than in face-to-face consultations. However, generalizing these conclusions to the Endocrinology specialty as a whole in a robust manner needs more data to further corroborate these results.

4.3 Neurology

In the Neurology medical specialty, two doctors present a statistically significant difference between *Patient%* distributions in face to face and virtual consultations and one doctor presents a tendency to have higher *Patient%* in virtual consultations. The summary statistics of all doctors are shown in Table 4.3 and Figure 4.3 shows the violin plots of each doctor distributions.

Doctor D1's Mann-Whitney U test results in a *p-value* of 0.18, meaning no statistically significant difference exists between both distributions. However, the low *p-value* indicates a tendency to have a

Table 4.3: Neurology results and summary statistics

Doctor	Med_f	Med_v	*p-value*	Cohen's *d*
D1	45.8%	48.3%	0.18	-
D2	54.9%	65.9%	$3.1e-3$	0.47
D8	32.8%	49.9%	$0.5e-5$	0.72
Joint	44.7%	57.7%	$2.1e-5$	-

higher *Patient%* in virtual consultations. When looking at Figure 4.3, we see that D1's virtual distribution values are slightly higher than the face-to-face distribution, even though the medians have a negligible difference between them of around 3%. Nonetheless, when taking all the results into account, we can say that D1 presents a tendency to look more at the patient during virtual consultations when compared to face-to-face consultations.

Doctor D2's Mann-Whitney U test results in a *p-value* of 0.0031, meaning that there is a statistically significant difference between the face-to-face and virtual distributions with the Cohen's *d* effect size of 0.47 indicating a medium size effect when changing consultation environments. The violin plot of D2, shown in Figure 4.3 shows the clear difference between distributions with the virtual distribution being higher than the face-to-face distribution. This is also confirmed by the medians in Table 4.3, with the virtual distribution median being 65.9%, a difference of around 10% when compared to the face-to-face distribution median of 54.9%. Therefore, doctor D2 looks more at the patient during virtual consultations than in face-to-face consultations.

Doctor D8's Mann-Whitney U test results in a *p-value* of 0.000005 and a Cohen's *d* effect size of 0.72, meaning that there is a statistically significant difference between the face-to-face and virtual distributions with a medium effect size. The violin plot of D8, shown in Figure 4.3 shows the clear difference between the face-to-face and virtual distributions, with the virtual distribution being clearly higher than the face-to-face distribution. Additionally, the medians difference of around a 17% further support this disparity, with the virtual distribution median being much higher than the face-to-face distribution, 49.9% and 32.8% respectively. Thus, doctor D8 looks significantly more at the patient during virtual consultations when compared to face-to-face consultations.

The Mann-Whitney U test on the joint data distribution results in a *p-value* of 0.000021, meaning that, generally, changing consultation environments has statistically significant effect for the three doctors. Looking at Table 4.3, we can further confirm that the median of the face-to-face distribution is indeed lower than the virtual distribution median, 57.7% and 44.7% respectively. Therefore, in general, the Neurology doctors have a tendency to look more at the patient during virtual consultations than in face-to-face consultations.

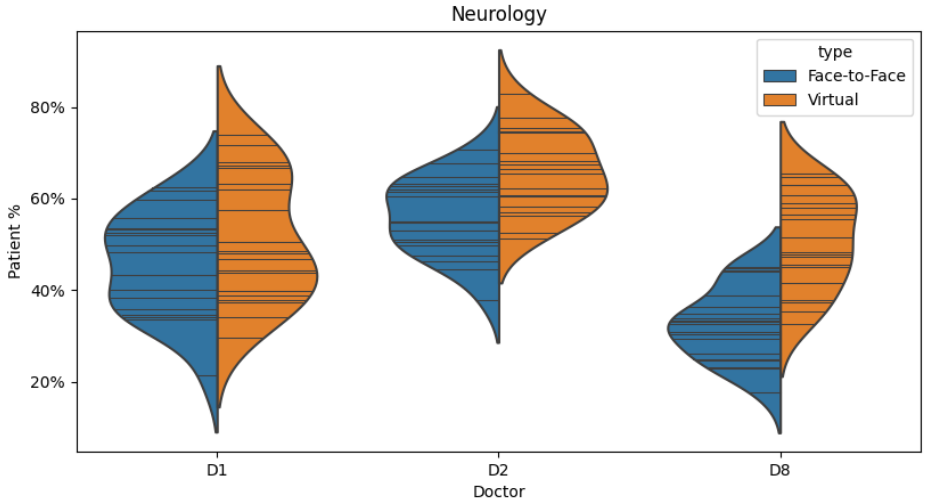


Figure 4.3: Neurology doctors violin plots

4.4 General and Family Medicine

In the General and Family Medicine medical specialty, two doctors present a statistically significant difference between *Patient%* distributions in face to face and virtual consultations and two doctors present a tendency to have higher *Patient%* in virtual consultations. The summary statistics of all doctors are shown in Table 4.4 and Figure 4.4 shows the violin plots of each doctor distributions.

Table 4.4: General and Family Medicine results and summary statistics

Doctor	Med_f	Med_v	*p-value*	Cohen's *d*
D4	38.2%	43.2%	0.049	0.31
D5	62.4%	59.3%	0.19	-
D7	46.4%	55.7%	0.11	-
D16	43.6%	59.2%	0.0059	0.54
Joint	46.5%	57.1%	$2.9e-4$	-

Doctor D4's Mann-Whitney U test results in a *p-value* of 0.049 with a Cohen's *d* size effect of 0.31, meaning there is a statistically significant difference between both types of consultations with a medium size effect. When looking at Figure 4.4, we can clearly see that D4's virtual distribution is higher than the face-to-face distribution. Additionally, the medians, shown in Table 4.4, confirm this with a difference of around 5%, with the virtual median being higher than the face-to-face median. For this reasons, we can say D4 spends more time looking at the patient during virtual consultations than in face-to-face consultations.

Doctor D5's Mann-Whitney U test results in a *p-value* of 0.19, meaning that there is no statistically significant difference between the face-to-face and virtual distributions. however the low *p-value* indicates a tendency for D5 to have a higher *Patient%* in the virtual consultations when compared to face-to-face consultations. This tendency is supported by the violin plot of D5's distributions, where we can see that

the virtual distribution extends further than the face-to-face distribution. However the medians of both distributions have a negligible difference between them of around 3%. Nonetheless, the *p-value* and the violin plots indicate a slight tendency for D5 to spends more time looking at the patient in virtual consultations.

Doctor D7's Mann-Whitney U test results in a *p-value* of 0.11, meaning that there is no statistically significant difference between the face-to-face and virtual distributions. Despite of that, both the violin plot and the medians indicate a tendency for D7 to have a higher *Patient%* in virtual consultations. The violin plot of D7 shows the peak of the virtual distribution being slightly higher than the face-to-face, in addition to the medians having a difference of around 9%. For this reasons, we can say that doctor D7 has a tendency to look more at the patient during virtual consultations than in face-to-face consultations.

Doctor D16's Mann-Whitney U test results in a *p-value* of 0.0059 with a Cohen's *d* size effect of 0.54, meaning there is a statistically significant difference between both types of consultations with a medium size size effect. When looking at the violin plot of D16's distributions, we can clearly see that the virtual distribution is higher than the face-to-face distribution. Additionally, the medians, shown in Table 4.4, confirm this with a difference of around 16%. Therefore, taking into consideration all results we can say D16 spends significantly more time looking at the patient during virtual consultations than in face-to-face consultations.

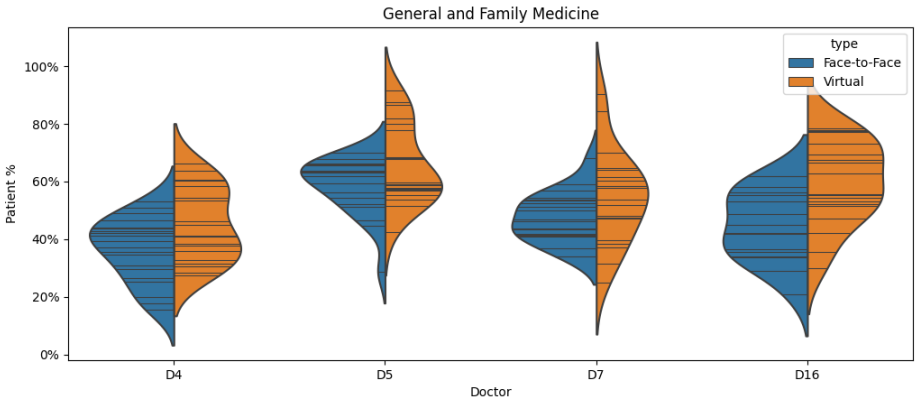


Figure 4.4: General and Family Medicine doctors violin plots

The Mann-Whitney U test on the joint data distribution results in a *p-value* of 0.00029, indicating the general tendency for doctors from this specialty to look more at the patient during virtual consultations. Looking at Table 4.4, we can further confirm that the median of the face-to-face distribution is indeed lower than the virtual distribution median, 46.5% and 57.1% respectively. Therefore, in general, the General and Family Medicine doctors have a clear tendency to look more at the patient during virtual consultations than in face-to-face consultations.

4.5 Consultation Heatmaps

In this section, we will show sample *PoG* heatmaps of every doctor. The heatmaps show us how the *PoG*s of each doctor were distributed along the screen plane, the darker the shade of blue the more the doctor looked at that zone of the plane. In addition to the heatmaps, the classification zones used are also shown, in many cases the zones used to classify the doctor's gaze needed to be adapted to the doctor being analyzed, since each doctor liked to organize their workspace a little differently. The different organizations of workspaces mainly consist in different keyboard positioning and computer screen positioning, more specifically the height and distance to the doctor. These different organizations led to the need of having a different divisions of the doctor's field of view. The heatmaps are divided in medical specialties, with each doctor presenting one face-to-face consultation heatmap and one virtual consultation heatmap.

In every medical specialty, we are able to discern the 3 main of interests of the doctor during a consultation: the patient, the screen and the keyboard. In almost all the cases we can see the clusters of darker shades of blue around these zones of interest, with exception of the keyboard zone which was the less looked at zone out of the 3 main zones. However, in various face-to-face heatmaps, we can also see that the cluster that is formed around the patient is not very concentrated and might cover a significant area. The two reasons for these effect consist in: the accuracy of the method under extreme head pose conditions in conjunction with the doctor wearing a mask and the patient changing positions during the consultation (adjusting his chair or posture). Due to these factors, in face-to-face consultations, the area covered by the patient's cluster is larger than the screen cluster due to the fluctuation of the gaze estimates. Nonetheless, in almost every case the patient cluster is discernible.

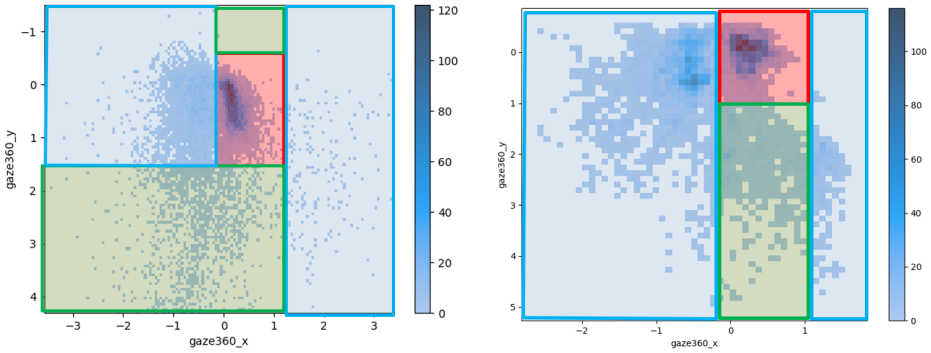
The screen cluster is also visible in every consultation and its very concentrated, indicating a tendency for the doctor to spend a large amount of its time looking at the screen.

The keyboard cluster position and size changes significantly depending on the doctor, due to the fact that each doctor has different habits for typing on the keyboard. Doctors D6, D12, D14, D1, D8, D3, D5 all orient their keyboard in the direction of the patient in face-to-face consultations, which increases the size of the keyboard classification zone in relation to the ideal one, as we can see for example in Figure 4.12a. Doctors D12, D14, D1, D10, D15, D16 all position their keyboard in the direction of the patient screen in virtual consultations, , as we can see for example in Figure 4.7b.

Overall, the patterns of clusters shown in the heatmaps are consistent throughout the 4 medical specialties. The existent differences between doctors' gaze patterns are mainly due to their personal preferences and characteristics, with no medical specialty having a specific type of pattern that distinguish from the others.

4.5.1 Gynaecology/Obstetrics

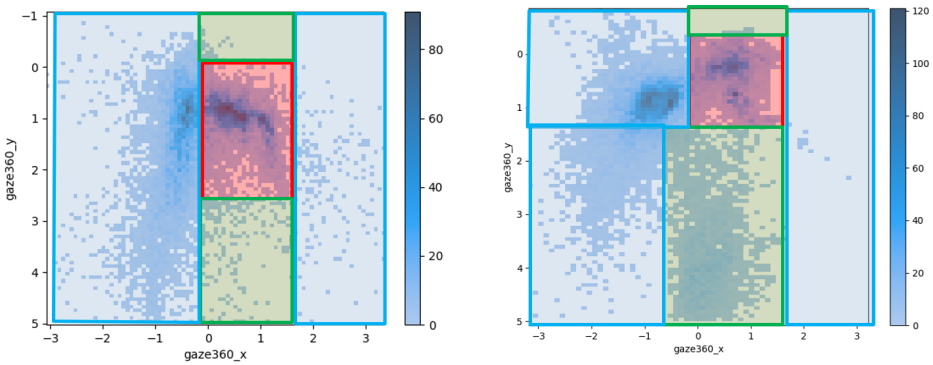
In this section, we show the *PoG* heatmaps of the 4 doctors from the Gynaecology/Obstetrics medical specialty.



(a) D6 face-to-face consultation

(b) D6 virtual consultation

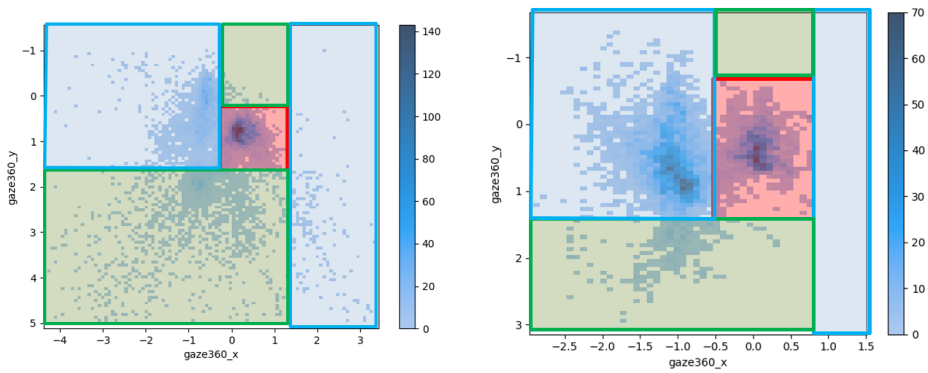
Figure 4.5: The *PoG* heatmaps of doctor D6



(a) D9 face-to-face consultation

(b) D9 virtual consultation

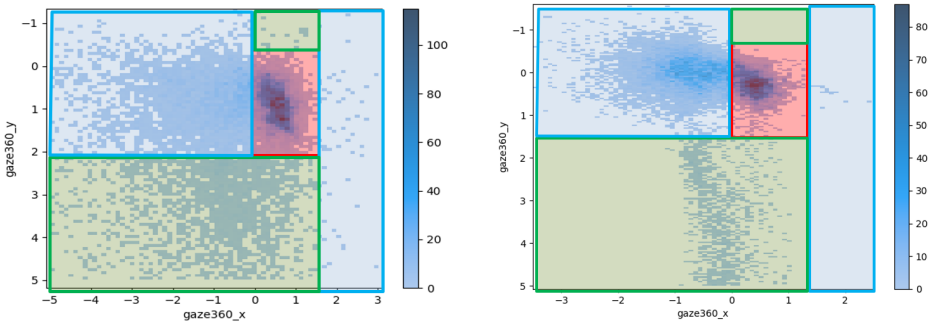
Figure 4.6: The *PoG* heatmaps of doctor D9



(a) D12 face-to-face consultation

(b) D12 virtual consultation

Figure 4.7: The *PoG* heatmaps of doctor D12



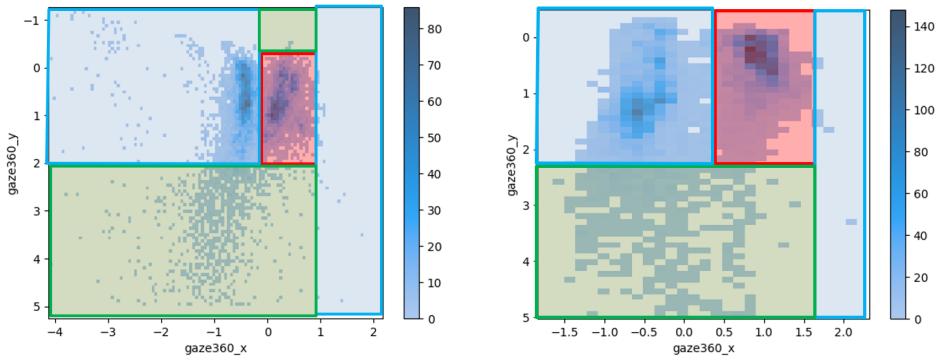
(a) D14 face-to-face consultation

(b) D14 virtual consultations

Figure 4.8: The *PoG* heatmaps of doctor D14

4.5.2 Neurology Heatmaps

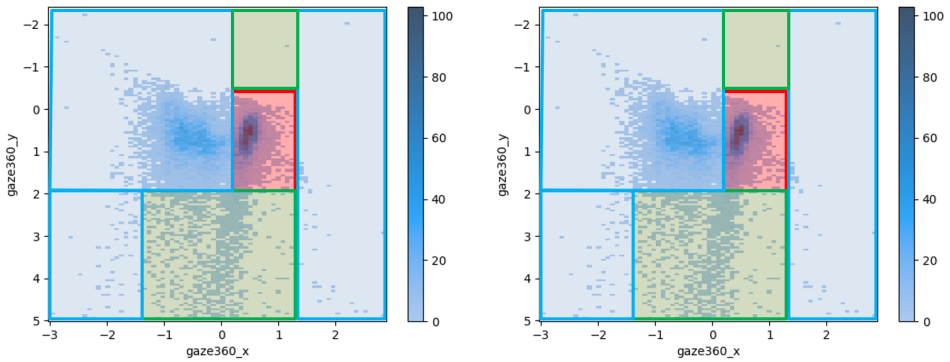
In this section, we show the *PoG* heatmaps of the 3 doctors from the Neurology medical specialty.



(a) D1 face-to-face consultation

(b) D1 virtual consultation

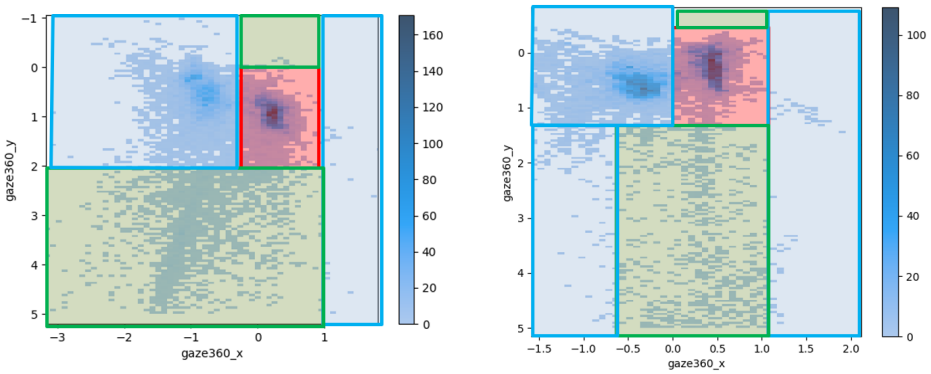
Figure 4.9: The *PoG* heatmaps of doctor D1



(a) D2 face-to-face consultation

(b) D2 virtual consultation

Figure 4.10: The *PoG* heatmaps of doctor D2



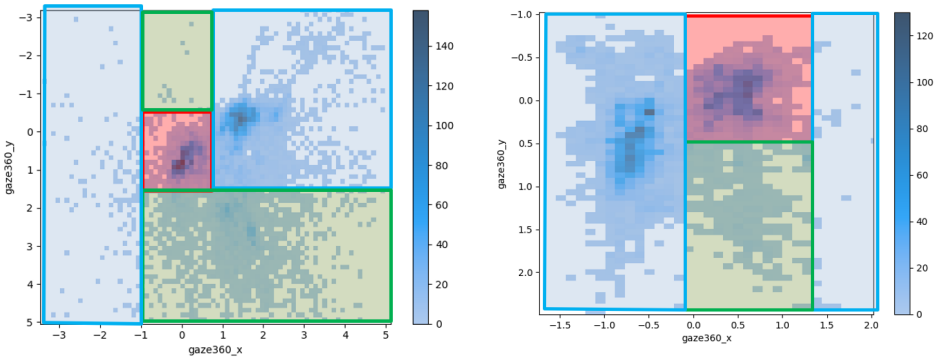
(a) D8 face-to-face consultation

(b) D8 virtual consultation

Figure 4.11: The *PoG* heatmaps of doctor D8

4.5.3 Endocrinology Heatmaps

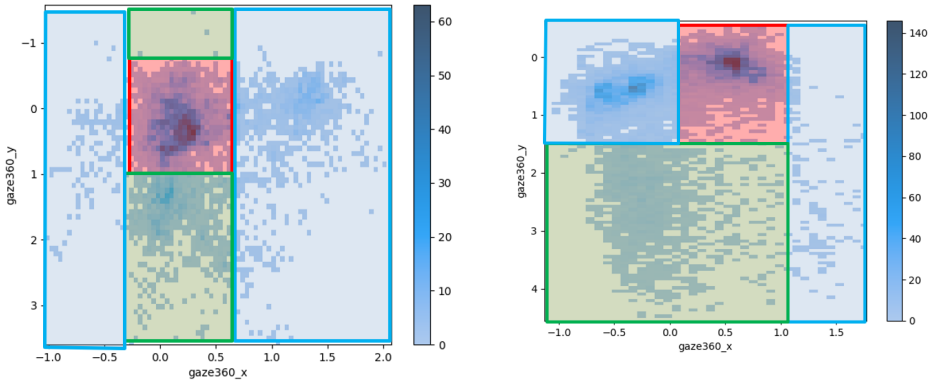
In this section, we show the *PoG* heatmaps of the 3 doctors from the Endocrinology medical specialty.



(a) D3 face-to-face consultation

(b) D3 virtual consultation

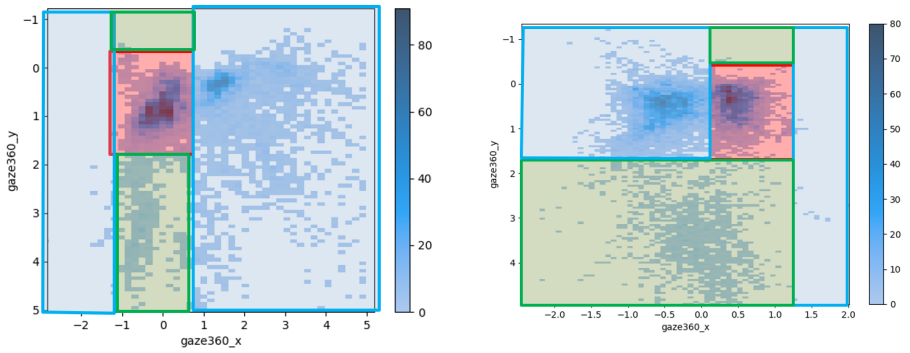
Figure 4.12: The *PoG* heatmaps of doctor D3



(a) D10 face-to-face consultation

(b) D10 virtual consultation

Figure 4.13: The *PoG* heatmaps of doctor D10



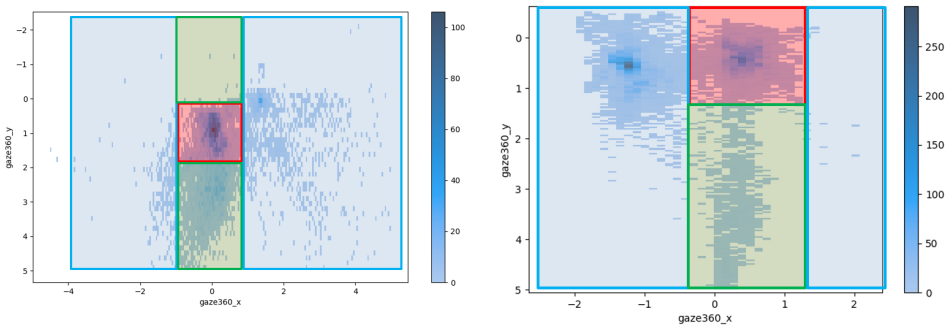
(a) D15 face-to-face consultation

(b) D15 virtual consultation

Figure 4.14: The *PoG* heatmaps of doctor D15

4.5.4 General and Familiar Medicine

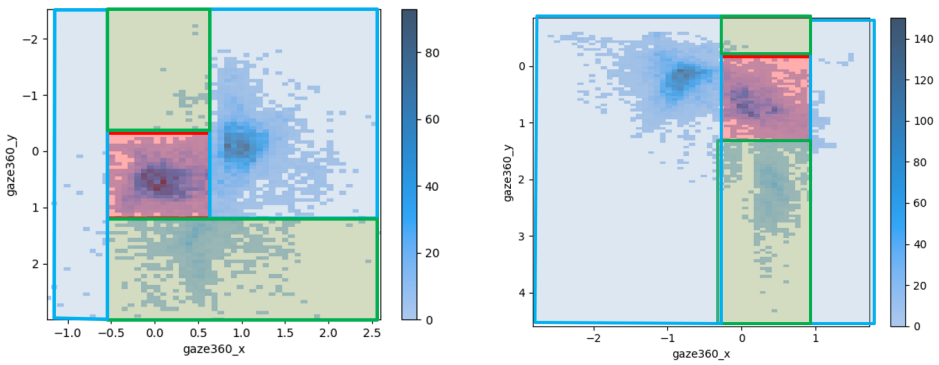
In this section, we show the *PoG* heatmaps of the 4 doctors from the General and Familiar Medicine medical specialty.



(a) D4 face-to-face consultation

(b) D4 virtual consultation

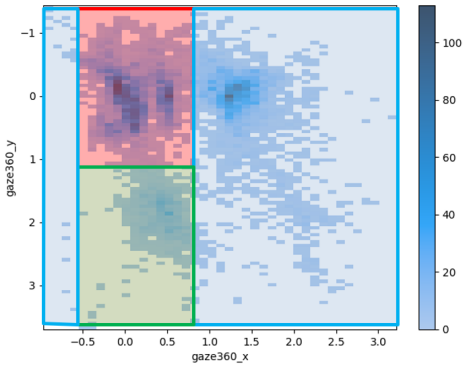
Figure 4.15: The *PoG* heatmaps of doctor D4



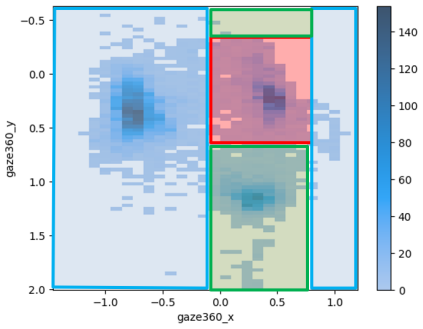
(a) D5 face-to-face consultation

(b) D5 virtual consultation

Figure 4.16: The *PoG* heatmaps of doctor D5

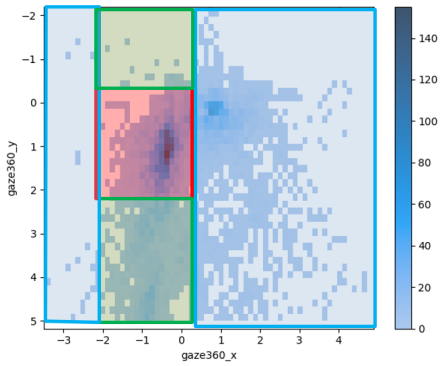


(a) D7 face-to-face consultation

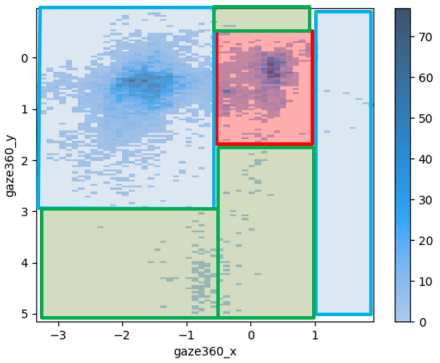


(b) D7 virtual consultation

Figure 4.17: The *PoG* heatmaps of doctor D7



(a) D16 face-to-face consultation



(b) D16 virtual consultation

Figure 4.18: The *PoG* heatmaps of doctor D16

4.6 Discussion

Overall our findings provide another insight into the doctor-patient relationship. More importantly, they provide a look at the effect of virtual consultations on the doctor-patient relationship, specifically on their effect on the doctor's behaviour. In three of the four medical specialties analyzed, we found a pronounced general tendency for doctors to look more at the patient during virtual consultations. However, the other medical specialty (Gynaecology/Obstetrics) seems to be the least affected by the consultation environment change. Half of the doctors present no differences or tendencies between consultation environments, leading to the Gynaecology/Obstetrics specialty presenting very few tendencies between consultation environments. Even though these results are not enough to make robust generalizations about medical specialties as a whole, they still provide an insight into the effect of virtual consultations on doctor behaviour. Consequently, these promising initial results deserve further study and corroboration to better understand the doctor-patient relationship.

Concerning the individual doctors' results, these can be analyzed robustly through the Mann-Whitney U test and Cohen's *d* effect size results. Out of the fourteen participating doctors, seven showed statistically significant differences with medium-sized size effects, while another five showed tendencies to

have higher *Patient%* in the virtual environment, with only two showing no differences in gaze behaviour. Overall, these results show the virtual consultation environment's clear effect on the doctors from the study population, since all of the doctors, except for two, are affected by the change in consultation environments. Additionally, these results also show the virtual consultation medium does not degrade eye contact since no doctor looked at the patient less during virtual consultations.

Our study has some limitations. The first limitation relates to a caregiver being present and the patient in some face-to-face consultations recorded. The way we classify gaze estimates classified any gaze towards the caregiver as gaze towards the patient, which might lead to some overestimation. However, the goal of measuring the division of the doctor's attention between screen and patient is still accomplished. The second limitation is related to the way we classify gaze direction, which leads to the *Patient%* metric being a estimation of the real time percentage the doctor spends looking at the patient. The zone used to define the *Patient* zone was the area on the side of the screen where the patient is located we can see that the *Patient* zone encapsulates more than just the patient's face. Thus, the *Patient* zone not only defines where the patient is located but also the surrounding area. The assumption used to make this division is that whenever the doctor looks to the side of the screen where the patient is, the doctor is looking at the patient. This is a strong assumption which leads to very good estimations of the real time the doctor spends looking at the patient through the *Patient%* metric. Another possible issue that was accounted for before the start of the study is related to the Hawthorne effect. The Hawthorne effect is a psychological effect that refers to the change in subjects' behaviours when they know they are being observed. To mitigate these effects, we kept the actions the doctor had to make that did not follow the regular consultation routine to a minimum, which is accomplished by only asking the doctor to click the *Record/Stop* button in the Recording interface in the beginning and at the end of the consultation.

Our study also has its strengths. The use of an automated gaze estimation pipeline led to the capture of a much larger number of interactions than previous studies of gaze in applied settings. While [26] provided 100 clinical interactions divided between 14 doctors, our study provided 260 face-to-face clinical interactions and 260 virtual clinical interactions. The use of an automated gaze estimation pipeline also allows us to study the gaze in more detail than with manually annotated gaze estimates. Additionally, using an appearance-based gaze estimation system like *Gaze360* provides a method that can be implemented in a variety of situations since it only needs an off-the-shelf webcam to be feasible.

In addition to assessing the virtual consultation environment's effect on doctor gaze behaviour, this study also provides, to our knowledge, the first implementation of an appearance-based gaze estimation pipeline in the clinical setup using an off-the-shelf webcam. This task came with many challenges, as mentioned in Section 3.3.1. In addition to the usual robustness against head pose, glasses and environment changes, the gaze estimation system had to be able to deal with the obligatory use of masks. Since the use of masks was not common before the pandemic, all the gaze estimation datasets and systems were not tailored for the task of estimating the gaze of a mask-wearing subject. The solution, which consisted in using a different mask that did not occlude the mouth and nose region, worked remarkably well through the various tests performed and continued to work throughout the study.

In the end, the gaze estimation pipeline implemented provides an easy way to record and analyze the doctor's gaze behaviour during consultations. It opens up many applications in the study of non-verbal communication in the doctor-patient relationship, some of which we will suggest in Section 5.1.

Chapter 5

Conclusions

In conclusion, this study successfully implemented a 3D appearance-based gaze estimation pipeline in the clinical setup. Additionally, the gaze estimation pipeline assessed the impact of the virtual consultations on the amount of time the doctors from the study population spend looking at the patient. It found that doctors from the study population spend significantly more time looking at the patient during virtual consultations, with a few exceptions. Also, when looking at the medical specialties analyzed, we found that three out of the four specialties present prominent tendencies to look more at the patient during virtual consultations when compared to face-to-face consultations. None of the specialties look less to the patients, suggesting that virtual consultations do not impact the eye contact between doctor and patient negatively. Conversely, these results show the sizable positive impact of virtual consultations on doctors' gaze behaviour.

The successful implementation of the gaze estimation pipeline allows us to analyze doctors' gaze behaviour in more detail and robustly than before. This ability is crucial to analyzing the impact of non-verbal communication in the physician-patient relationship to improve the quality and efficiency of the health care provided. The applications of gaze estimation in the clinical setup are immense. It could be used in the assessment of clinical setup changes. We assessed the impact of virtual consultations against face-to-face consultations. However, many other setup changes can also be analyzed. Software updates can be analyzed to increase doctors' efficiency when navigating the interfaces on the computer. Consultation room changes impact on doctor's gaze behaviours like lighting conditions changes and patient positioning changes can be analyzed thoroughly. More interestingly, gaze estimation can be used for digital biomarker analysis. During a consultation, the patient's attention can be analyzed to see how it evolves throughout the visit. Children's gazes during pediatric therapy sessions could be analyzed to assess if they are focusing on what the therapist wants and improve the overall therapy sessions in the future. Like pediatric therapy, patients with neurological/psychological disorders gaze could be studied during therapy sessions to assess the effectiveness of the therapy sessions and possibly improve them in the future.

In summary, doors opened by gaze estimation are immense, which, coupled with the study of other non-verbal cues, will enable the study of the relationship between people's non-verbal behaviours and

actions.

5.1 Future Works

The impact of virtual consultations on doctors' gaze behaviour is apparent. However, it should continue to be analyzed to provide us with a better understanding of the consequences to the physician-patient relationship. Therefore, a large scale study involving more doctors with a higher number of consultations recorded should be performed. A study like this is essential to support and corroborate the initial results obtained to fully understand the impact of virtual consultations on doctors' gaze behaviour. In addition to a larger amount of samples collected, this future project could also include the extraction of more gaze features along with the *PoG*s, in the form of saccades, smooth pursuits or even face landmarks. This would allow for a deeper study of the doctor's gaze patterns.

Another work that could be done in support of this study is on the correlation between the *Patient%* metric and the actual percentage of time the doctor spent looking at the patient. To obtain the actual percentage, the doctor has to put on the eye-trackers that look like glasses, which provide images of the Field-Of-View (FOV) of the doctor and the gaze located in the FOV image (method used in [26]). Finding the exact relationship between the *Patient%* and the actual percentage of time is an important step on the support of these studies results.

Additionally, we need to study more than gaze to study the total impact of non-verbal communication in the physician-patient relationship. Therefore, coupling gaze estimation with other types of machine learning algorithms to analyze other non-verbal cues like voice pitch, monotony, facial expressions, or body posture is the next step in the studying the impacts of non-verbal communication in the physician-patient relationship.

Bibliography

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *Proceedings of the International Conference on Engineering and Technology (ICET)*, 2017.

[2] K. Alberto Funes Mora and J.-M. Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2018.

[4] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. *Advances in Neural Information Processing Systems*, 1993.

[5] R. S. Beck, R. Daughtridge, and P. D. Sloane. Physician-patient communication in the primary care office: a systematic review. *The Journal of the American Board of Family Medicine*, 2002.

[6] J. Bird and S. Cohen-Cole. The three-function model of the medical interview. an educational device. *Advances in psychosomatic medicine*, 1990.

[7] J.-Y. Bouguet. Camera calibration toolbox for matlab. 2001.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] Y. Cheng, H. Wang, Y. Bao, and F. Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *CoRR*, 2021.

[10] J. Cohen. Statistical power analysis jbr the behavioral. *Sciences. Hillsdale (NJ): Lawrence Erlbaum Associates*, 1988.

[11] T. Fischer, H. J. Chang, and Y. Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[12] K. Funes Mora, F. Monay, and J.-M. Odobez. Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ACM Symposium on Eye Tracking Research and Applications (ETRA)*, 2014.

[13] A. George and A. Routray. Real-time eye gaze direction classification using convolutional neural network. *CoRR*, 2016.

[14] A. Graves, S. Fernández, and J. Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, 2005.

[15] E. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 2006.

[16] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[17] D. Gutstein, E. Montague, J. Furst, and D. Raicu. Optical flow, positioning, and eye coordination: Automating the annotation of physician-patient interactions. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019.

[18] J. F. Ha and N. Longnecker. Doctor-patient communication: a review. *Ochsner Journal*, 2010.

[19] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, 1995.

[20] Y. Hart, E. Czerniak, O. Karnieli-Miller, A. E. Mayo, A. Ziv, A. Biegon, A. Citron, and U. Alon. Automated video analysis of non-verbal communication in a medical setting. *Frontiers in Psychology*, 2016.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV))*, 2015.

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[23] Q. Huang, A. Veeraraghavan, and A. Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 2017.

[24] L. A. Jeni and J. F. Cohn. Person-independent 3d gaze estimation using face frontalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Q. Ji and X. Yang. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 2002.

[26] C. Jongerius, H. Boorn, T. Callemein, N. Boeske, J. Romijn, E. Smets, and M. Hillen. Eye-tracking analyses of physician face gaze patterns in consultations. *Scientific Reports*, 2021.

[27] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[28] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, , and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[29] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[32] H. S. Lepper, L. R. Martin, and M. R. DiMatteo. A model of nonverbal exchange in physician-patient expectations for patient involvement. *Journal of Nonverbal Behavior*, 1995.

[33] G. Liu, Y. Yu, K. A. F. Mora, and J. Odobez. A differential approach for gaze estimation. *CoRR*, 2019.

[34] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[35] F. Lundh. An introduction to tkinter. *URL: www. pythonware. com/library/tkinter/introduction/index. htm*, 1999.

[36] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 1947.

[37] M. S. Mast. On the importance of nonverbal communication in the physician–patient interaction. *Patient Education and Counseling*, 2007.

[38] M. S. Mast and G. Cousin. The role of nonverbal communication in medical interactions: Empirical results, theoretical bases, and methodological issues. *The oxford handbook of health communication, behavior change and treatment adherence*, 2013.

[39] C. Morimoto and M. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 2005.

[40] E. Olson. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[41] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[42] S. Park, E. Aksan, X. Zhang, and O. Hilliges. Towards end-to-end video-based eye-tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[43] W. Rawat and Z. Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 2017.

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2015.

[45] B. Smith, Q. Yin, S. Feiner, and S. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2013.

[46] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[47] K. Takahashi, S. Nobuhara, and T. Matsuyama. A new mirror-based extrinsic camera calibration using an orthogonality constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[48] T. Tan, E. Montague, J. Furst, and D. Raicu. Robust physician gaze prediction using a deep learning approach. In *Proceedings of the IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020.

[49] S. Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006, 2006.

[50] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 2012.

[51] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the sˆ 3gp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[52] L. R. Young and D. Sheena. Survey of eye movement recording methods. *Behavior research methods & instrumentation*, 1975.

[53] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.

[54] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[55] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, 2017.

[56] X. Zhang, Y. Sugano, and A. Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2019.

[57] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[58] Z. Zhu and Q. Ji. Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on Biomedical Engineering*, 2007.

[59] P. H. Zimmerman, J. E. Bolhuis, A. Willemsen, E. S. Meyer, and L. P. Noldus. The observer xt: A tool for the integration and synchronization of multimodal signals. *Behavior research methods*, 2009.

Appendix A

Informed Consent Form

HOSPITAL DA LUZ

Consentimento para Tratamento de Dados Pessoais no Âmbito de Estudo/Ensaio Clínico

Nome completo _____

Nº Processo _____ Data Nascimento _____/_____/_____

Se disponível, colar neste campo a etiqueta de identificação do cliente.

Nome do Estudo / Ensaio clínico _____

Centro do Estudo / Ensaio clínico _____

Morada do Centro do Estudo / Ensaio Clínico _____

Este ato pressupõe a prévia obtenção de autorização para participação no estudo/ensaio clínico acima indicado, formalizada em consentimento informado assinado pelo Participante e/ou seu Representante Legal.

Parte declarativa do Investigador Principal

Confirmo que informei o participante acima identificado, enquanto titular dos dados objeto do presente consentimento, e/ou o seu representante legal (se aplicável), de forma completa, inteligível e adequada à sua capacidade de compreensão, sobre o âmbito, o objeto e a finalidade da utilização dos dados do participante, incluindo imagens (designadamente de exames complementares de diagnóstico, de procedimentos médicos e/ou cirúrgicos, vídeos e fotografias), obtidos no âmbito do estudo/ensaio clínico acima indicado, garantindo que os seus dados recolhidos nesse contexto serão tratados de forma confidencial apenas no âmbito supra identificado e o seu acesso reservado à equipa de investigação. Foi também explicado que os seus dados, depois de anonimizados ou pseudonimizados, podem ser partilhados com terceiros para quem pode ser transferida a propriedade dos direitos autorais das imagens e que esses dados podem ser utilizados futuramente em trabalhos científicos, artigos de revisão, publicações médicas, livros didáticos, apresentações congressos, cursos e outras finalidades de âmbito científico ainda que diferentes da acima indicada, destinadas apenas a profissionais de saúde. Em complemento dos esclarecimentos prestados foi disponibilizada a Política de Privacidade desta unidade de saúde que é a Responsável pelo Tratamento dos Dados.

Respondi a todas as questões que me foram colocadas e assegurei-me de que foi compreendida toda a informação fornecida oralmente e por escrito, da ausência de dúvidas por esclarecer e de que houve um período de reflexão suficiente para a tomada da decisão. Informei ainda de que este consentimento (se concedido) poderá ser retirado a todo o tempo, clarificando que a retirada do consentimento não compromete a licitude do tratamento de dados que foi feito com base no mesmo até ao momento em que seja retirado. Também garanti que, em caso de recusa deste consentimento (ou de ser futuramente retirado), não existirá qualquer efeito sobre os cuidados de saúde prestados, podendo, no entanto, ficar prejudicada a possibilidade de participação no estudo/ensaio clínico.

Nome do Investigador Principal: _____

☐ Médico ☐ Outro Profissional de Saúde (categoria / função): _____

Identificação nº: _____ ☐ Cédula Profissional ☐ Nº Mecanográfico

Contacto institucional do Investigador Principal: _____

Assinatura do Investigador Principal: _____ Data: _____/_____/_____

Ao Participante titular dos dados / Representante Legal (de participante menor ou incapaz de prestar consentimento)

Por favor, leia com atenção todo o conteúdo deste modelo e da informação complementar disponibilizada (se aplicável). Solicite mais informações se tiver dúvidas ou não estiver completamente esclarecido. Verifique se todas as informações estão corretas. Só assine este modelo se tudo estiver conforme, se estiver totalmente esclarecido e se concordar com o seu conteúdo integral. Assine-o perante o Investigador Principal e conserve um exemplar na sua posse.

Fui previamente informado e esclarecido, em linguagem clara e acessível, acerca da utilização para fins científicos dos meus dados obtidos no âmbito do estudo/ensaio clínico acima indicado, incluindo apresentações, trabalhos científicos e publicação de casos clínicos tratados nesta unidade e/ou nas unidades integradas no grupo Luz Saúde. Foi-me garantido que os meus dados serão tratados de forma confidencial, sendo o seu acesso exclusivamente reservado à equipa de investigação que está obrigada a proteger a minha privacidade e a não divulgar qualquer informação pessoal que não seja estritamente necessária no âmbito do estudo/ensaio clínico acima indicado e nos termos da legislação aplicável. Foi-me também explicado que os meus dados podem ser partilhados com terceiros, desde que sejam **previamente anonimizados ou pseudonimizados** para que não exista hipótese de ser individualmente identificado o seu titular através da informação

Nota: Assinado em suporte digital ou em suporte físico digitalizado e anexado ao processo clínico; entregar ao cliente o original do documento físico ou o documento digital impresso ou disponibilizá-lo no Portal do Cliente.

contida nesses dados. Nessas circunstâncias, os referidos dados devidamente anonimizados ou pseudonimizados poderão ser utilizados futuramente em trabalhos científicos, artigos de revisão, publicações médicas, livros didáticos, apresentações, congressos, cursos e outras finalidades de âmbito científico ainda que diferentes do estudo/ensaio clínico acima indicado, destinadas apenas a médicos e a outros profissionais de saúde. Estes trabalhos científicos, apresentações e afins têm como objetivo divulgar informação que pode ser vir a ser útil à comunidade científica. Habitualmente, este tipo de trabalhos aborda casos que, pela sua originalidade ou raridade de apresentação, abordagem ou resultados atingidos, merecem ser divulgados, com a finalidade de contribuir para o desenvolvimento do conhecimento médico/científico e a melhoria dos cuidados de saúde prestados a outros clientes. Embora se procure garantir o total anonimato, fui alertado que pode haver algum risco (muito limitado) de que eu próprio ou outra(s) pessoa(s) possa(m) reconhecer-me.

Apreendi toda a informação que me foi disponibilizada oralmente e por escrito. Tive prévio conhecimento da Política de Privacidade da unidade, em conformidade com a legislação aplicável. Dessa forma estou perfeitamente informado sobre a identidade do Responsável pelo Tratamento e os termos do tratamento dos meus dados, incluindo o fundamento e a finalidade, bem como prazos de conservação, os contactos do Encarregado da Proteção de Dados e o modo como poderei exercer os meus direitos. Foi-me dada a oportunidade de fazer todas as perguntas sobre o assunto e para todas elas obter resposta esclarecedora, garantindo-me que não haverá prejuízo para os meus direitos assistenciais se recusar esta solicitação, apenas com a ressalva de que a recusa ou retirada deste consentimento poderá prejudicar a possibilidade de participar no estudo/ensaio clínico em referência. Concedo este consentimento, enquanto fundamento de legitimidade para o tratamento dos meus dados no âmbito do estudo/ensaio clínico acima indicado, como uma contribuição voluntária e gratuita no melhor interesse do desenvolvimento científico, da investigação médica e da formação e ensino, sem direito a qualquer retribuição ou compensação de qualquer espécie. Tive o tempo suficiente para refletir sobre a proposta e para tomar a decisão que aqui expresso de forma consciente, livre, informada e voluntária. Estou ciente de que poderei retirar o seu consentimento a todo o tempo, compreendendo que a retirada do consentimento não compromete a licitude do tratamento de dados que foi feito com base no mesmo até ser retirado.

Declaro que autorizo o tratamento dos meus dados obtidos no âmbito do estudo / ensaio clínico acima identificado. Concedo que os meus dados só sejam utilizados para qualquer outra finalidade nos termos acima indicados, se previamente anonimizados ou pseudonimizados.

Assinatura do Participante (com 16 anos ou mais e com discernimento)**:**

_____ **Data:** _____ /_____ /_____

Representante Legal (de Participante menor de 18 anos ou com incapacidade de prestar este consentimento)

A decisão do Representante Legal deve refletir a vontade presumível do Participante titular dos dados.

Nome completo: _____

Contacto: _____

Tipo de Representação: ☐ **Poder Paternal** ou ☐ **Tutor** ou ☐ **Curador** ou ☐ **Acompanhante nomeado pelo Tribunal**

ou ☐ **Procurador de Cuidados de Saúde, com poderes para praticar este ato**

Identificação nº: _____ ☐ Cartão de Cidadão ☐ Bilhete de Identidade ☐ Passaporte

☐ Outro: _____

Assinatura do Representante Legal: _____ Data: _____ /_____ /_____

Nota: Assinado em suporte digital ou em suporte físico digitalizado e anexado ao processo clínico, entregar ao cliente o original do documento físico ou o documento digital impresso ou disponibilizá-lo no Portal do Cliente.