Master Thesis

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Neural Network Based Eyetracking System

Bc. Adrián Pitoňák

Supervisor: Ing. Jaromír Doležal, Ph.D.
Field of study: Open Informatics
May 2022

MASTER'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Pitoňák Adrián** Personal ID number: **474380**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Specialisation: **Computer Vision and Image Processing**

II. Master's thesis details

Master's thesis title in English:

Neural Network Based Eyetracking System

Master's thesis title in Czech:

Systém sledování očních pohybů založený na neuronových sítích

Guidelines:

1) Review the state of the art methods used in eyetracking.
2) Propose an appropritate system architecture with respect to real-time operation.
3) Implement selected methods and train the system on an dataset of annotated images.
4) Experimentally evaluate speed and accuracy of detection.

Bibliography / sources:

[1] Lin, Horng-Horng, et al. "Pupil localization for ophthalmic diagnosis using anchor ellipse regression." 2019 16th International Conference on Machine Vision Applications (MVA). IEEE (2019)
[2] Vera-Olmos, Francisco Javier, et al. "DeepEye: Deep convolutional network for pupil detection in real environments." Integrated Computer-Aided Engineering 26.1 (2019)
[3] Fuhl, Wolfgang, et al. "BORE: Boosted-oriented edge optimization for robust, real time remote pupil center detection." Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (2018)
[4] Prasad, Dilip K., and Maylor KH Leung. "Clustering of ellipses based on their distinctiveness: An aid to ellipse detection algorithms." 2010 3rd International Conference on Computer Science and Information Technology. Vol. 8. IEEE (2010)
[5] Eivazi, Shaharam, et al. "Improving real-time cnn-based pupil detection through domain-specific data augmentation." Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (2019)

Name and workplace of master's thesis supervisor:

Ing. Jaromír Doležal, Ph.D. CIIRC ČVUT v Praze

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **10.12.2021** Deadline for master's thesis submission: **20.05.2022**

Assignment valid until: **30.09.2023**

Ing. Jaromír Doležal, Ph.D. Supervisor's signature	prof. Ing. Tomáš Svoboda, Ph.D. Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt Student's signature

Acknowledgements

First of all, I wish to express my sincere thanks to my supervisor Ing. Jaromír Doleºal, Ph.D. for his enthusiasm for the project, for his support, encouragement and patience. Further, I would like to thank the faculty, for providing me with all the necessary facilities for the research. To conclude, I cannot forget to thank my family and close friends for all the unconditional support throughout the study.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, 13. May 2022

Abstract

The goal of this thesis is to create a framework for training eye-tracking models. Before the creation of the framework, a comprehensive analysis of methods used for this task was undergone. The work solves a part of the problem using neural networks designed for regression and the other part of the problem using common image processing techniques. Several neural networks were considered, including networks designed for this task but also networks designed for a general regression or classi cation task. The created framework integrates novel libraries, neural network architectures and methods. The framework is subsequently tested on the dataset of real subjects. Within the testing, the work compares several integrated neural network architectures in terms of their accuracy and speed. Finally, the chosen models are exported to a standard format and inserted in a simple video pipeline that runs in real-time.

Keywords: eye-tracking, pupil detection, neural networks, image processing

Abstrakt

Cie©om tejto práce je vytvori´ framework na trénovanie modelov pre sledovanie o£ných pohybov. Pred vytvorením samotného frameworku prebehla rozsiahla analýza metód pouºívaných na túto úlohu. Práca rie²i £as´ úlohy pomocou neurónových sietí ur£ených na regresiu a druhú £as´ pomocou známych metód na spracovanie obrazu. V rámci regresie, boli vzaté do úvahy siete, ktoré sú priamo ur£ené na túto úlohu, ale aj v²eobecné siete ur£ené pre regresiu a klasi káciu. Vytvorený framework integruje novodobé kniºnice, architektúry a metódy neurónových sietí. Framework je následne testovaný na datasete reálnych subjektov. V rámci testovania práca porovnáva rôzne architektúry neurónových sietí, ktoré boli vo frameworku integrované, na základe ich presnosti a rýchlosti. Na záver sú vybrané modely exportované do ²tandardného formátu a zasadené do jednoduchého systému pre spracovanie obrázkov, ktorý beºí v reálnom £ase.

Klí£ová slova: sledování o£ních pohyb·, detekce pupily, neuronové sít¥, spracování obrazu

P°eklad názvu: Systém sledování o£ních pohyb· zaloºený na neuronových sítích

Contents

Chapter 1

Introduction

Eye-tracking has found its application in a great variety of fields. In medicine, it is an essential tool to assist surgeries, such as laser eye surgery [MEK+01]. It also serves to identify abnormalities in vision and regions of interest of patients with mental diseases such as schizophrenia, bipolar disorder, Alzheimer's disease, and others [HK18]. In psychology, the main fields of application are unquestionably "visual search, reading, natural tasks, scene viewing and other information processing" [MF12]. In marketing, it provides valuable data on eye movements and regions of interest to select and refine products [WP08]. In the virtual reality and gaming industry, eye tracking helps in rendering complex scenes at high frame rates by defining the field of view of a user [CKK19]. It is without a question that one would find many other applications of eye-tracking, especially in recent years.

The task of eye-tracking is to estimate where one is looking, i.e. the gaze vector. Advancements in both software and hardware in recent years allow for non-invasive, real-time and affordable solutions to this task. Thanks to their accessibility, eye-tracking systems gained a great amount of popularity in both the academic and commercial sectors in recent times. Many affordable eye-tracking systems are based on video oculography [KKK+16] [KSM+19a] [WRvA+19]. The gaze vector can be estimated either directly from video frames (images) or indirectly by extracting features from the image and calculating the gaze vector using a geometrical model of the eye. The core of such an indirect feature-based model is an image processing pipeline that takes an image as input and outputs the features required for gaze vector estimation.

This thesis has two main goals: to create a framework for an eye-tracking image processing pipeline that integrates state of the art libraries and to test it on a real-world dataset provided by the supervisor of this thesis by comparing integrated neural network architectures. The final models should work in real-time in an optimized image processing pipeline. The thesis is structured as follows. In Chapter 2, video-oculography eye-tracking methods are introduced. After that, some of the state of the art feature extraction models are depicted. In Chapter 3, the aim of the thesis is described in more detail. Chapter 4 describes the framework built for this work, including integrated libraries, neural network architectures and the process of training them. Models that were chosen to be employed in the final pipeline are explained afterwards. The chapter ends with a description of how these models were evaluated. Consequently, the results are presented in Chapter 5. The results are discussed afterwards in Chapter 6, and suggestions for future improvements are provided. Finally, in Chapter 7, the achieved results of the whole work are summarized.

Figure 1.1: (a) Video oculography eye-tracking system. The user is looking at the monitor while the camera is sensing his eye movements. Optionally, there's an external infrared illumination reflected from the eye.

Chapter 2

State of the art

Increased interest and demand for affordable eye-tracking systems induced a great wave of research into video oculography models. The next section introduces the video oculography model in great detail. Such models can be classified into two main types [HJ09] [LCM19], namely appearance-based and feature-based methods. As this thesis focuses on feature-based methods due to their better accuracy, appearance-based methods are only outlined. After introducing both methods the NvGaze [KSM+19a] neural network, which was implemented in the framework of this work and chosen for the final models, is described comprehensively.

2.1 Video oculography

Video oculography is a non-invasive method for measuring eye movements. Fundamental parts of an eye-tracking system based on video oculography are a screen and a video camera, optionally accompanied by external infrared illumination (see Figure 1.1). These methods are classified into two main categories [HJ09] [LCM19]: appearance-based methods [TKA02] [SK10] [KKK+16] [KSM+19a] [WRvA+19] and feature-based (also called model-based) methods [LWP05] [YAR+19] [HEM+20]. In general, both methods work as follows: an image is fed to the model, processed through some pipeline that outputs a gaze vector. The following sections cover specifics of the gaze vector retrieval and its representation.

2.1.1 Appearance-based methods

Appearance-based methods [TKA02] [SK10] [KKK+16] [KSM+19a] [WRvA+19] only use a camera to capture an image of eyes and rely solely on the appearance of the eye [HJ09]. The model learns a mapping function from an image directly to a gaze vector. The gaze vector is usually represented as a point on the screen or the rotation angles of the eye regarding the head position [LCM19]. In other words, these models are trained end-to-end and unlike feature-based methods, they do not need any geometrical eye model (see Figure 2.1).

These methods work great under bad light conditions and with low-resolution cameras, where higher robustness is traded for lower accuracy [LCM19]. That is why they found their application in mobile devices and other inexpensive setups [KKK+16]. Best results were achieved using convolutional neural networks [LB+95]. To achieve reasonable accuracy they require a large dataset, which can be either real-world [KKK+16] [KSM+19a] [ZPB+20] or synthetic [KSM+19a]. Convolutional neural network models are capable of achieving real-time performance [KSM+19a] [WRvA+19].

Figure 2.1: Appearance-based methods general principle (from Krafka et al. [KKK+16]) Schema of neural network GazeCaputre [KKK+16] demonstrating end-to-end training of appearance based methods.

2.1.2 Feature-based methods

Feature-based methods consist of two parts: a feature extraction model and a geometrical model. A feature extraction model is trained to extract local features such as contours, eye corners, and eye reflections of the external light sources (further referred to as glints) [LCM19] (see Figure 2.2). Features are then passed to a geometrical model that outputs a gaze vector. The gaze vector is usually represented as a 2D point on the screen (further referred to as the point of gaze) as in the case of appearance-based methods or a 3D direction vector from which a point of gaze can be calculated. This depends on the geometrical model. Here are the main characteristics of both approaches.

Figure 2.2: Feature-based methods general principle. Eye features such as pupil, cornea and corneal reflections (glints) are extracted from the image. These features are passed to a geometrical model that estimates the gaze vector.

3D approach

The 3D model-based approach yields more robust results but requires a more complex setup in terms of knowing eye parameters that cannot be directly measured and camera parameters [LCM19]. Some of the eye-trackers that use this approach are DeepVOG [YAR+19], RemoteEye [HEM+20] and MagicEyes [WRvA+20]. The 3D approach has two parts. First, locate eye features, i.e. pupil centres and glints. Second, estimate the gaze vector. The first part of this approach is covered in Section 2.2. As for the second part, a geometrical 3D model was developed by Guestrin et al. [GE06]. This model requires at least one camera and at least two external light sources when the head is not stationary and also a subject-specific calibration procedure. More advanced models by Swirski et al. [SD13] and Dierkes et al. [DKB18] overcome these requirements. In addition, all of the methods require that light sources and monitor positions are known when the system is not head-mounted to estimate camera translation and rotation matrices.

As the method by Geusterin et al. [GE06] introduces parameters of the eye that are essential for gaze vector reconstruction, it is described in this section in great detail. After that, the other above mentioned methods are outlined. The geometrical model developed by Guestrin et al. [GE06] for one camera and several external light sources can be derived in the following way (see Eq. 2.1 to 2.14 and Figure 2.3). The derivation is reprinted from [GE06].

Figure 2.3: 3D geometric eye model (from Guestrin et al. [GE06]).

Assuming point light sources and pinhole cameras a ray that comes from the light source i, l_i, and reflects at the point q_{ij} on the corneal surface, modeled as a convex spherical mirror of radius R, such that the reflected ray passes through the camera centre j, o_j, and intersects the camera image plane at the point u_{ij}, following equations can be formulated.

$$q_{ij} = o_j + k_{q,ij}(o_j - u_{ij}) \text{ for some } k_{q,ij} \quad (2.1)$$

$$\|q_{ij} - c\| = R \quad (2.2)$$

The law of reflection indicates that points l_i, q_{ij}, c and o_j are co-planar and angles $\angle\, l_i q_{ij} c$ and $\angle\, o_j q_{ij} c$ are equal:

$$(l_i - o_j) \times (q_{ij} - o_j) \cdot (c - o_j) = 0 \quad (2.3)$$

$$(l_i - q_{ij}) \cdot (q_{ij} - c) \cdot \|o_j - q_{ij}\| = (o_j - q_{ij}) \cdot (q_{ij} - c) \cdot \|l_i - q_{ij}\| \quad (2.4)$$

Similarly, considering a ray that comes from pupil centre p, refracts at the point r_j on the corneal surface, passes through the camera centre j, o_j, and intersects the camera image plane at the point v_j, following equations can be obtained.

$$r_j = o_j \quad k_{r;j} (o_j \quad v_j) \text{ for some } k_{r;j} \tag{2.5}$$

$$jj r_j \quad cjj = R \tag{2.6}$$

The law of refraction indicates that points p, r_j, c and o_j are co-planar and from Snell's law (i.e. $n_1 \sin \; _1 = n_2 \sin \; _2$) following equations can be formulated.

$$(r_j \quad o_j) \quad (c \quad o_j) \quad (p \quad o_j) = 0 \tag{2.7}$$

$$n_1 \;\; jj(r_j \quad c) \quad (p \quad r_j)jj \;\; jj\, o_j \quad r_j\, jj = n_2 \;\; jj(r_j \quad c) \quad (o_j \quad r_j)jj \;\; jj\, p \quad r_j\, jj \tag{2.8}$$

where n_2 is the index of refraction of air (1).

Finally, considering the distance K between the pupil centre and the centre of corneal curvature leads to

$$jj p \quad cjj = K \tag{2.9}$$

If eye parameters (R, K and n_1) are obtained through calibration, substituting Eq. 2.1 into Eq. 2.3 yields the following system of equations.

$$(l_i \quad o_j) \quad (u_{ij} \quad o_j) \quad (c \quad o_j) = 0 \tag{2.10}$$

Assuming single camera is used (i.e. j = 1), the system can be rewritten in matrix from as

$$\underbrace{\begin{bmatrix} [(l_1 \quad o) \quad (u_1 \quad o)]^T \\ [(l_2 \quad o) \quad (u_2 \quad o)]^T \\ \vdots \\ [(l_N \quad o) \quad (u_N \quad o)]^T \end{bmatrix}}_{M} (c \quad o) = 0 \tag{2.11}$$

From the co-planar constraint (see Eq. 2.10) it follows that matrix M has at most rank 2 and the solution has the following form.

$$c - o = k_{c,b} b_{norm} \tag{2.12}$$

When $N = 2$ the system can be solved in the following manner.

$$b_{norm} = \frac{b}{||b||} \tag{2.13}$$

$$b = [(l_1 - o) \times (u_1 - o)] \times [(l_2 - o) \times (u_2 - o)] \tag{2.14}$$

Which is together with calibration parameters sufficient to reconstruct the optic axis $p - c$ of the eye.

The downside of this model is that eye parameters (R, K and n_1) must be estimated through several points calibration procedure. The model by Swirski et al. [SD13] solves this problem and requires only a sequence of images from a single camera to estimate necessary parameters, removing also the necessity of external illumination. The core operation of this model is "unprojection" of pupil ellipses from each image into a circle in 3D space, considering full perspective projection. For the full explanation and optimization of the method, see [SD13]. This method was further improved by Dierkes et al. [DKB18] using a more realistic eye model [LG68].

For all of these methods, if the point of gaze is requested, the visual axis needs to be calculated from the optic axis. For doing this, subject-specific parameters have to be estimated, typically from the calibration procedure, in which the subject is required to fixate on a single point [GE06] or these parameters can be approximated by anthropomorphic averages [SD13]. After that, visual axes of both eyes need to be transformed from the image coordinate system to the world coordinate system. To do so, intrinsic and extrinsic camera parameters must be obtained, typically through a camera calibration procedure. For a detailed description of the point of gaze retrieval, see [GE06].

2D approach

The 2D approach is much simpler in terms of initial setup, as no additional information about the position of eye-tracking system parts nor camera calibration are required. This approach includes subject-specific eye parameters and camera parameters directly into a mapping function from 2D feature space like Pupil-Centre-Corneal-Reflections, contours, etc. to the point of gaze [LCM19]. A mapping function can be determined generally, so that no user specific calibration is required, but this approach suffers from lower accuracy caused by head motions. As a result, recent methods employ per subject or even per experiment calibration typically with neural network as a mapping function [WZS16] [HHM+18]. A popular method based on algorithmic approach is Starburst [LWP05].

2.2 Eye feature extraction models

As mentioned earlier, feature-based methods are based on finding distinctive features of the eye to employ them in a geometrical data model. Common features for eye localization include pupil, cornea and corneal reflections, also known as glints. Some notable state of the art algorithms are based on an algorithmic approach [SFK18] [FEH+18] [FGS+18]. The other state of the art models are based on neural networks. Currently, two types of neural networks are commonly used for this task, specifically segmentation networks [YAR+19] [CKA+19] [KCB+21] and regression networks [FSK+17] [KSM+19a] [ESK+19].

The biggest advantage of models based on an algorithmic approach is that they run on CPU within a few milliseconds [FGS+18]. The spectrum of methods employed in these models is very broad. Most commonly used are Canny edge detector [Can86], e.g. [KPB14] [FKS+15] [FSKK16] [SFK18], Haar-like features [VJ01], e.g. [BD12], Gabor filters, e.g. [TKC00] [CK07]. These methods usually deal with an ellipse fitting problem that can be solved, for example, with the RANSAC algorithm [FB81] as in [LWP05] [‘BD12]. According to NvGaze Paper [KSM+19a], which compares most of the methods mentioned above on the PupilNet dataset [FSK+17], Circular binary features algorithm [FGS+18] or CBF was the only method that yielded similar results as the neural network based NvGaze and outperformed it in terms of precision when pixel error threshold for detection was lower (see Figure 2.4). Besides high accuracy, CBF runs at a very short inference time of 0.44 milliseconds to 6.8 milliseconds on a CPU, depending on the resolution and approach, which is lower or very similar in comparison to other models [FGS+18].

Figure 2.4: Comparison of state of the art feature extractors (from Kim et al. [KSM+19a]). Original description, where "We, Our and Ours" refer to NvGaze model [KSM+19a]: "Average pupil estimation error on PupilNet datasets. Top: We compare the average detection rate of our pupil estimation network against Starburst [LWP05], Set [JHB+15], Swirski et al. [‘BD12], ExCuSe [FKS+15], ElSe [FSKK16], PupilNet [FSK+17], Park et al. [PZBH18] and CBF [FGS+18]. Bottom: The 5-pixel error is averaged across individual PupilNet datasets [FSK+17] (bold marker) and bounded by best and worst error values for all datasets (upper and lower markers). Our approach reaches highest robustness. Note that for CBF only the average detection rates over all datasets were published, not the detection rate for individual datasets." The metrics is described as follows, "Accuracy of pupil estimation is usually given in form of a probability of estimating the pupil location with a maximum distance of 5 pixels from the ground truth pupil location (5-pixel error or "detection rate")" [FSK+17].

On the other hand, the use of neural networks for this task becomes more and more popular lately [GSS+19]. In short, neural networks, as the core of deep learning, process input vector, e.g. an image, through several layers of pre-trained weights and output a prediction vector. The weights are trained through learning procedure called backpropagation [RHW86]. As previously stated, there are two main categories of neural networks. The first one treats the problem as a segmentation problem. Neural networks in this category employ Unet-like [RFB15] architecture that consists of an