Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

YIZHAN HUANG

Neural Network Eye Tracking:Determining Nine-Grid Regions & Application

DEGREE PROGRAMME IN DATA ENGINEERING
2023

Author(s) HUANG, YIZHAN	Type of Publication Bachelor's thesis	Date 06/2023
	Number of pages 24	Language of publication: English

Title of publication
Neural Network Eye Tracking:Determining Nine-Grid Regions & Application

Degree Programme
Artificial Intelligence

Abstract

As technology advances, the development of mobile platforms has nearly reached saturation. AR/MR (Augmented or Mixed Reality) eyewear is anticipated to be the new type of product leading technological advancements in the next 5-10 years. Some AR/MR glasses on the market with specialized functions have already become lightweight and convenient to use. Combining AR/MR glasses with eye-tracking technology implies that the latest technological advancements will significantly impact the daily life interactions and communications of users who are unable to use conventional interaction modes. Compared to the existing eye-tracking technology in AR/MR glasses, such as infrared eye-tracking, employing machine learning will make this technology more robust and adaptive to different environments, as well as providing better compatibility during calibration and setup. This thesis will demonstrate the use of cameras combined with neural networks for eye-tracking analysis and application scenarios, and hypothesize on applying this technology to future AR/MR devices that will be compatible with this technology.

Keywords

AR,MR,eye-tracking,interaction methods,mahcine learning,nerural Network

CONTENTS

1 INTRODUCTION

With the advancement of society, there is an increasing prevalence of individuals encountering varying degrees of mobility constraints. These constraints can be attributed to a plethora of causes including vehicular accidents, natural degradation of physical capabilities due to aging, or congenital disorders impeding the typical range of motion anticipated in individuals. Consequently, this results in a loss of conventional abilities to interact with humans and objects. Examples of such conditions include Amyotrophic Lateral Sclerosis (ALS), spinal cord injuries due to trauma, neurological impairments, or congenital muscular dystrophy among others (Cipresso et al.,2011, pp.320-324). Studies indicate (Koskas & Héran,2013) that these ailments share a commonality; depending on the severity, patients may lose varying degrees of motor and verbal capabilities, but retain a certain level of ocular mobility which may also diminish as the condition progresses. With technological advancements, contemporary Augmented Reality/Mixed Reality (AR/MR) eyewear platforms offer the potential for novel modalities of interaction via eye tracking for individuals who are immobilized or possess limited mobility throughout their lifespan.

The application of eye-tracking through machine learning is not confined to medicinal purposes. In contexts of quotidian use or work-related scenarios, compared to current eye-tracking technologies on AR eyewear devices, such as infrared eye-tracking, the incorporation of machine learning renders this technology more robust and adaptable to varied environments (Holmqvist et al.,2013a, pp.45-52). It also offers superior compatibility in terms of calibration and configurations (Zhang et al.,2015, pp.4511-4520). Hence, the adoption of this technology for eye-tracking may signify a trend in the foreseeable future.

Therefore, this thesis presents an implementation framework for eye-tracking using machine learning. In Chapter 2 provides a detailed exposition on the methodologies for data collection and the specifications of the equipment utilized in the data collection environment. In Chapter 3 delves into a comprehensive analysis of the selection and architecture of the machine learning models, accompanied by a comparative analysis with results from pre-trained models. Lastly, In Chapter 4 shows the prospective directions and future developments of this technology.

2 PLANNING AND IMPLEMENTATION OF EXPERIMENTAL DATA COLLECTION

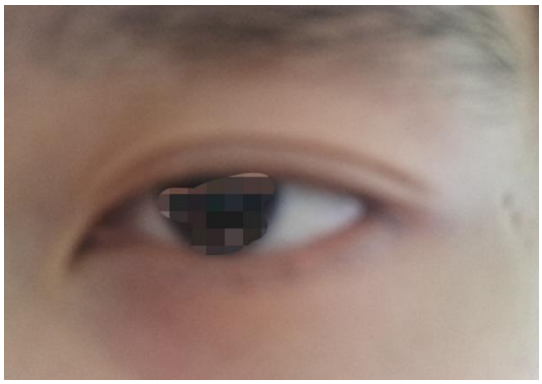
2.1 Information on Data Collection Equipment and Collection Protocol:

To ensure an appropriate level of generalization in the model after training, this study involves collecting ocular data from individuals of diverse ethnicities, with stringent protection of the participants' privacy. The data will be exclusively used for research purposes and not for commercial applications. Due to the limitations of the available hardware, MR devices will not be utilized for data collection and experimentation; instead, OAK-1 and Dell monitors (model unspecified), as well as an RTX2080 graphics card and an i7-9700K CPU, will be used for conventional data collection, model training, and testing. Participants must be free of eye diseases. Additionally, owing to the constraints of the test environment and hardware, and in accordance with the ocular conjugacy mentioned in (Robinson,1964), human eyes exhibit identical movement paths in the absence of eye diseases (excluding myopia). As such, this study will employ a monocular image collection approach. Specific experimental standards are provided in appendix 1 of this document, and the data collection and testing processes will be detailed in subsequent article.

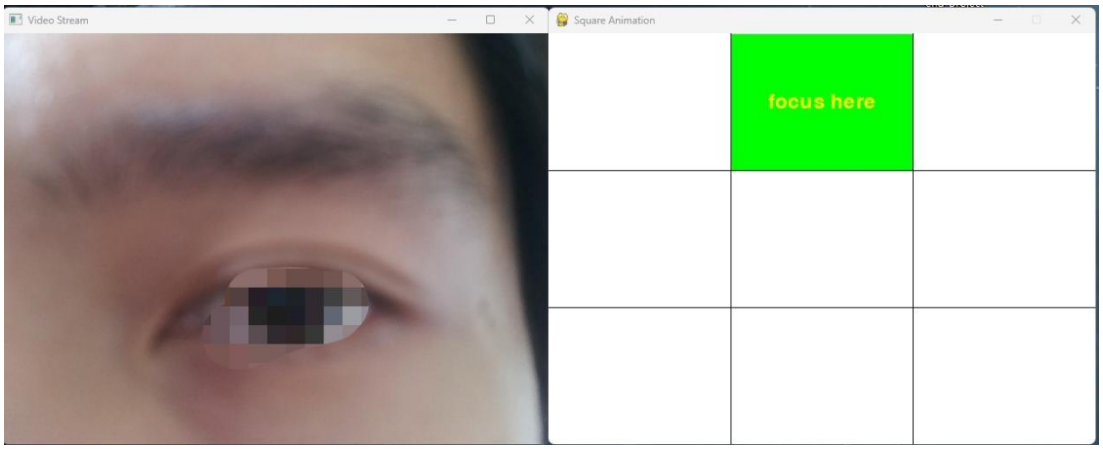
2.2 Experiment Data Collection Procedure:

This experiment involved a total of seven participants. The data collection and testing environment are illustrated in [Figure 1]. A 3x3 grid was created within a 640x480 window. Following a countdown, one of the squares within the grid would randomly illuminate in green for two seconds and then turn off, followed by a one-second pause before the next cycle. Participants were instructed to continuously focus on the green square throughout the process. This cycle was repeated 50 times, with the entire data collection procedure, including the countdown, taking approximately 4 minutes and 50 seconds. Two windows were displayed on the screen during the experiment; one showing the grid, and another showing a live feed from the camera, allowing participants to monitor in real-time and ensure data validity refer to [Figure 3]. Examples of the images collected are displayed in [Figure 2]. Considering potential changes in equipment and shooting conditions, there were no specific clarity requirements for the photographs as long as the contour of the eyes and the positions of the pupils and sclera were clearly discernible, similar to employing Gaussian blur. This will not impact the model's final determinations and may enhance its generalization capabilities. Images were saved in JPG format with a resolution of 640x480.

Figure[1]-Collection processing



Figure[2]-Example of data



Figure[3]-Example of monitoring processing

3 MACHINE LEARNING - MODEL, DATA CONSTRUCTION, AND TRAINING PROCESS

3.1 Final Target:

The ultimate goal of this training is to capture the human eye in real-time through a camera, and utilize the trained model to track and determine the specific grid position the subject is currently gazing at, subsequently illuminating it in the test program. Thus, this is fundamentally a classification problem, with ten classification targets corresponding to the nine positions in a 3x3 grid, as well as the determination of closed eyes or invalid images.

3.2 Reason of Model Selection & Previous Studies introduction:

Historically, Convolutional Neural Networks (CNNs) have demonstrated remarkable efficacy in image classification tasks, particularly within the realm of artificial intelligence (Hiojazi et al.,2015). For instance, in the study presented in (Yang et al.,2017,pp. 533-54), a two-stage Deep Convolutional Neural Network (DCNN) was employed for the automatic analysis of Diabetic Retinopathy (DR). DCNN, an extension of CNN, in this neural architecture, a local network initially extracts lesion-specific features from retinal images, classifying them into four categories: normal, microaneurysms, hemorrhages, and exudates. Subsequently, a global network holistically analyzes these features, grading the severity of DR based on international clinical DR grading standards. These grades encompass: no visible lesions or abnormalities, mild NPDR with only microaneurysms, moderate NPDR with abundant microaneurysms, hemorrhages, and hard exudates, and severe NPDR with venous abnormalities, large blot hemorrhages, cotton wool spots, venous beading, venous loops, and venous reduplications. To enhance grading accuracy, the algorithm incorporated an imbalanced attention mechanism on the input image, utilizing a weighted lesion map to boost the performance of the DR grading network. This methodology not only precisely detects lesions in retinal images but also evaluates the severity of DR. Experimentally, the eye model achieved an accuracy rate of 95.84% in detecting eye closure, while the face model reached an accuracy of 80.01%, showcasing the efficiency and precision of employing DCNN.

Furthermore, atricle (Alparslan et al.,2020) also shows the utilization of deep learning techniques for detecting driver fatigue was discussed. Given that drowsy driving is a leading cause of traffic accidents, the authors proposed a framework based on camera-captured frames to detect the degree of eye closure as a metric for assessing driver fatigue. To achieve this, two distinct datasets were used: the "Eye-blink" dataset specifically for eyes and the "Closed Eyes in the Wild (CEW)" dataset containing facial images. Based on these datasets, two models were developed. The first, termed the "eye model," was dedicated to detecting the degree of eye closure,

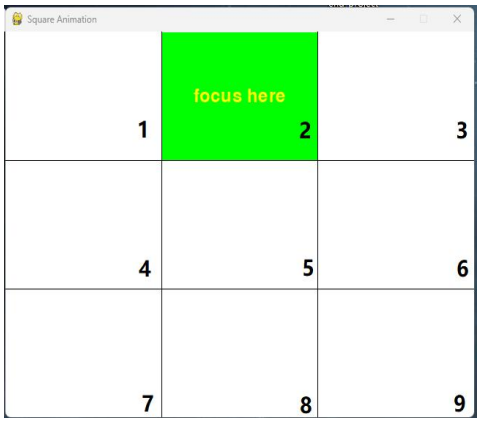
while the second, the "face model," was designed to detect the state of the entire face. Both models were constructed using a four-layer CNN. Experimentally, the eye model achieved an accuracy rate of 95.84% in detecting eye closure, while the face model reached an accuracy of 80.01%. Interestingly, the authors also discovered that adversarial training and data augmentation techniques could further enhance the accuracy of the face model.

Therefore, drawing on previous research, a CNN is employed in this study, constructed through PyTorch, to recognize and classify eye images. However, it is noteworthy that in 2022, Geoffrey Hinton introduced a novel neural network paradigm termed the Forward-Forward Algorithm (Hinton,G.,2022a). In his publication, Hinton posited that this algorithm could potentially offer commendable performance even under the constraints of limited computational resources. While this thesis does not delve into the specifics of the network, the Forward-Forward Algorithm emerges as a promising avenue for applications in AR/MR glasses, proffering a low-power and efficacious algorithm. Additionally, at the end of Section 3, a pre-trained ResNet-18 model will be juxtaposed with the custom-built model for a comparative analysis of their performance and results.

3.3 Model Data and Network Architecture:

3.3.1 Model Data Architecture:

Throughout the experiment, a total of 1489 images focused on the eye region were collected, of which 1219 images were allocated to the training and validation sets, and 270 images were designated for the test set. Among the 1219 images, 75 percent were used for the training set, and 25 percent were employed for the validation set. Given the relatively small number of images available for training and testing, the model architecture was kept simple to prevent overfitting. The collected image dataset was annotated with labels ranging from 1 to 9 during the data collection process, in accordance with the program settings, as illustrated in [Figure 4]. The camera captured images of the participants' eyes when the corresponding grid square turned green and continued until the square turned off. Upon completion of the collection process, a CSV file was generated, which contained the image paths and corresponding square IDs for organizing and preparing the training set. Before consolidating all the image data, manual screening was conducted to annotate images of closed eyes [Figure 5] with Label 0, enabling the model to recognize the closed-eye state.



Figure[4]-Example of program response Figure[5]-Example of Label 0 picture

3.3.2 Model Network Architechture:

The current convolutional neural network (CNN) consists of two convolutional layers, a dropout layer with a dropout probability of 0.5, one max-pooling layer, and three fully connected layers, where the last fully connected layer is used for outputting the final 10 classes. The output channel sizes for the two convolutional layers are 16 and 32, respectively, with a kernel size of 5 for both, indicating that both convolutional layers have two 5x5 filters of the same size. Full visualization of CNN can be found in appendix 2

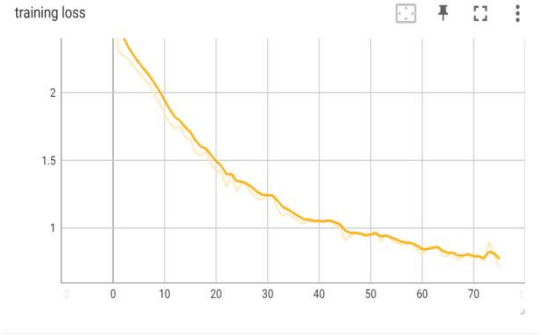
3.3.3 Model forward propagation pipeline:

During the forward propagation process, images first pass through two convolutional layers, each followed by a ReLU activation function and a max-pooling layer. The ReLU activation functions introduce non-linearity to the network, while the max-pooling layers reduce the spatial dimensions of the feature maps, which helps decrease computational load and capture larger spatial features. After convolution and pooling, the feature maps are flattened into one-dimensional vectors and processed through three fully connected layers for classification and regression. Dropout layers are applied after the first and second fully connected layers, which randomly deactivate neurons with a certain probability, serving as a regularization technique to prevent overfitting. Ultimately, the network outputs a vector containing ten values representing the probabilities of the image belonging to each of the ten categories. The category with the highest probability is selected as the predicted class for the image. Detailed network architecture and forward propagation pipeline will be thoroughly illustrated in APPENDIX 2.

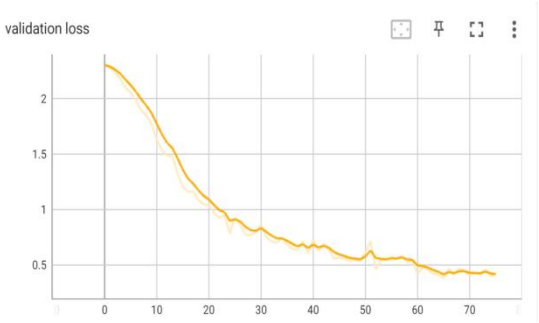
3.4 Presentation and comparison of the results of the self-built CNN model with the RetNet18 pre-trained model:

3.4.1 Display of CNN Model Results:

The training process of the CNN model is shown in Figures [6,7]. Through these figures, it can be observed that the model correctly converges and learns the features of the images during the training process. CNN modle cost 1.029 hours to meet early stop condition with 75 trainning steps and final training loss was approximately 0.71, and the final validation loss was approximately 0.41.The model was tested with 270 untrained images, with the results shown in the confusion matrix in Figure [10].



Figure[6]-traing loss of CNN



Figure[7]-validation loss of CNN

3.4.2 Presentation and Introduction of ResNet18 Model Results:

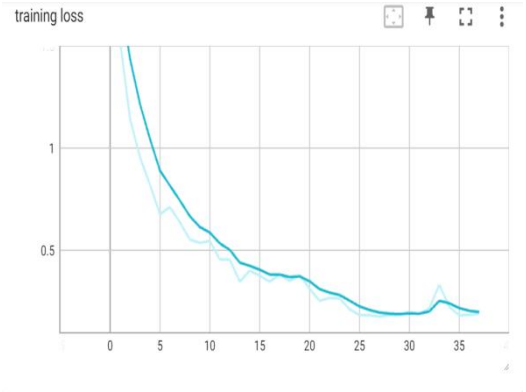
3.4.2.1 What is ResNet18:

ResNet18 is a deep residual neural network, first proposed in the article (He et al.,2016,pp.770-778). "18" represents the number of layers in this model. The unique feature of the deep residual network (ResNet) is its use of the concept of "residual learning" to address the issues of "vanishing gradients" and "network degradation" in

deep neural networks. Both of these issues can affect the performance of the network and even lead to a decrease in accuracy as the depth of the network increases. In traditional CNN networks, the output of each layer is computed based on the output of its previous layer. However, in ResNet, the design of the network allows the output of each layer to depend not only on the output of its previous layer but also to include some outputs from earlier layers. This forms what is known as "skip connections" or "shortcut connections". This design enables the network to directly learn the "residual" mapping between the input and output, and during the backpropagation process, the gradient can be propagated directly through these "shortcut connections", thereby addressing the problem of vanishing gradients. ResNet18, as a configuration of ResNet, is made up of 18 layers including convolutional layers, fully connected layers, and non-linear activation layers. Although its structure is relatively simple and its computational cost is relatively low, ResNet18 still performs exceptionally well due to its incorporation of the characteristics of residual learning. Therefore, for application scenarios with limited computational resources, ResNet18 becomes an excellent choice.

3.4.2.2 ResNet Results Display:

The training process of the ResNet18 model is shown in Figures [8,9]. Through these figures, we can see that the model correctly converges and learns the features of the images during the training process, similar to the CNN model. Resnet18 modle cost 9.213 minutes to meet early stop condition with 38 trainning steps and final training loss was approximately 0.18, and the final validation loss was approximately 0.41. The model was tested using 270 untrained images, with the results shown in the confusion matrix in Figure [11].



Figure[8]-traing loss of ResNet

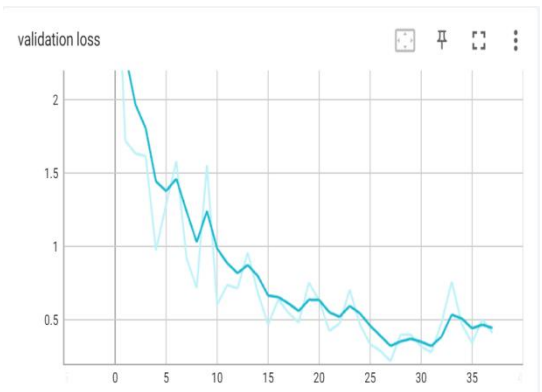


Figure [9]-validation loss of ResNet

3.4.3 Comparision between CNN and ResNet Results:

Despite numerous meticulous adjustments made to the parameters of both models, it is undeniable that the outcomes are still far from satisfactory. As discussed in article (Holmqvist et al.,2013b, pp.45-52), the quality and quantity of data play a decisive role in the efficacy of model training. The custom CNN model achieved an average accuracy of only 20.74% on the test set. As can be observed from the confusion matrix in Figure [10], this model exhibited higher numbers of classifications for images with labels 0, 8, and 4, with the highest accuracy rates recorded for labels 0 (80.77%), 6 (40%), and 4 (30.77%). Table [1] provides a detailed representation of the model's performance on the 10 labels in the test set.

Label	**Total Images**	**Correct Predictions**	**Accuracy (%)**	**Recall**	**Precision**
0	26	21	80.77	0.81	0.34
1	28	0	0.00	0.00	0.00
2	27	7	25.93	0.26	0.54
3	28	2	7.14	0.07	0.07
4	26	8	30.77	0.31	0.14
5	29	0	0.00	0.00	0.00
6	25	10	40.00	0.40	0.43
7	27	3	11.11	0.11	0.23
8	27	5	18.52	0.19	0.08
9	27	0	0.00	0.00	0.00

Table[1]-CNN result

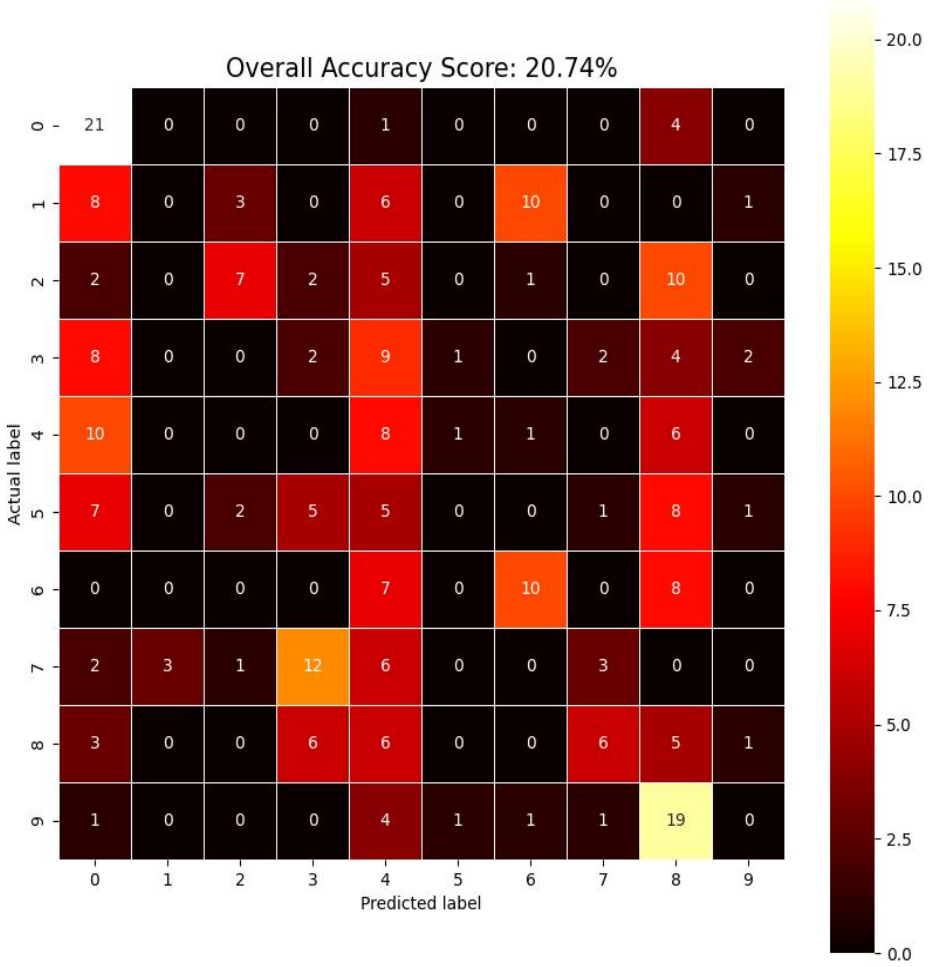
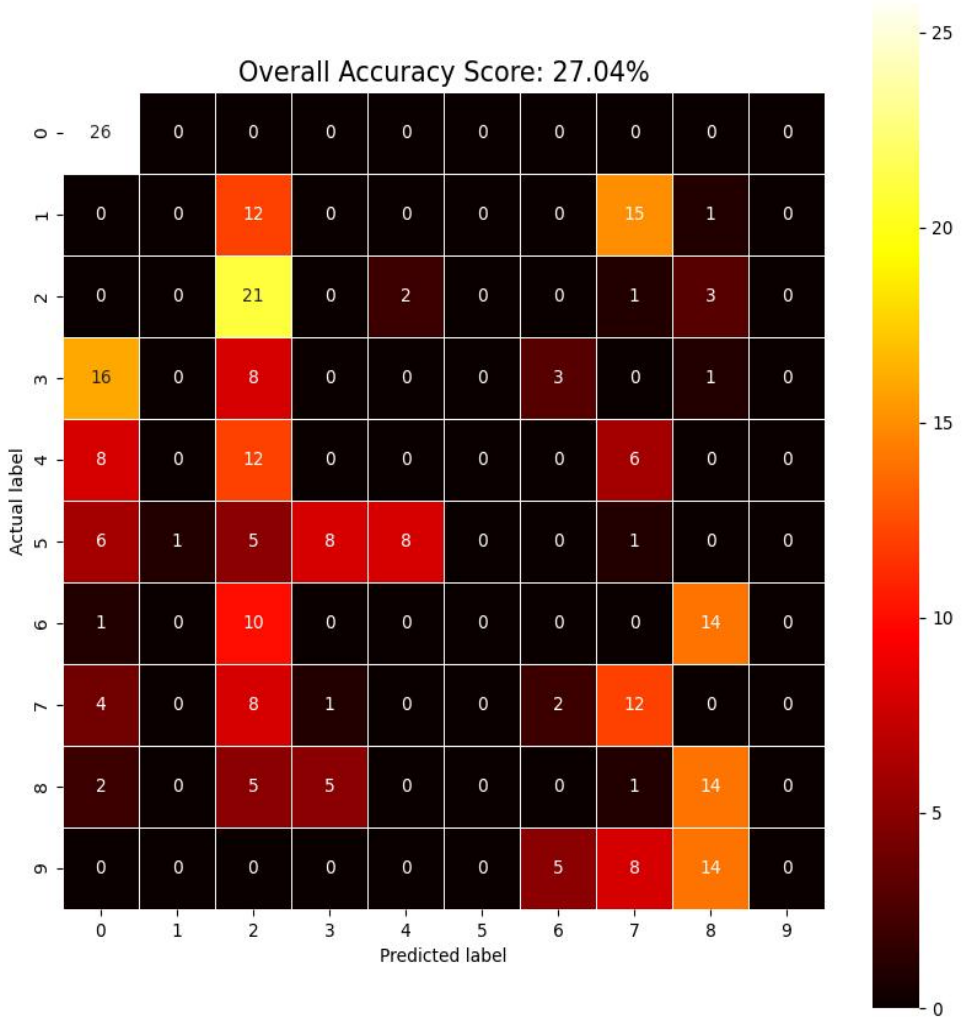


Figure [10]-CNN Confuse Martix

The ResNet model, however, achieved a higher average accuracy rate of 27.04% on the test set, approximately 7 percentage points higher than that of the CNN model. Similarly, from the confusion matrix in Figure [11], the ResNet model demonstrated higher numbers of classifications for images with labels 2, 0, and 8, with the highest accuracy rates achieved for labels 0 (100%), 2 (77.78%), and 8 (51.85%). Table [2] provides a comprehensive presentation of the model's accuracy rates for images across all labels.

Label	Total Images	Correct Predictions	Accuracy (%)	Recall	Precision
0	26	26	100.00	1.00	0.41
1	28	0	0.00	0.00	0.00
2	27	21	77.78	0.78	0.26
3	28	0	0.00	0.00	0.00
4	26	0	0.00	0.00	0.00
5	29	0	0.00	0.00	0.00
6	25	0	0.00	0.00	0.00
7	27	12	44.44	0.44	0.27
8	27	14	51.85	0.52	0.30
9	27	0	0.00	0.00	0.00

Table[2]-ResNet result



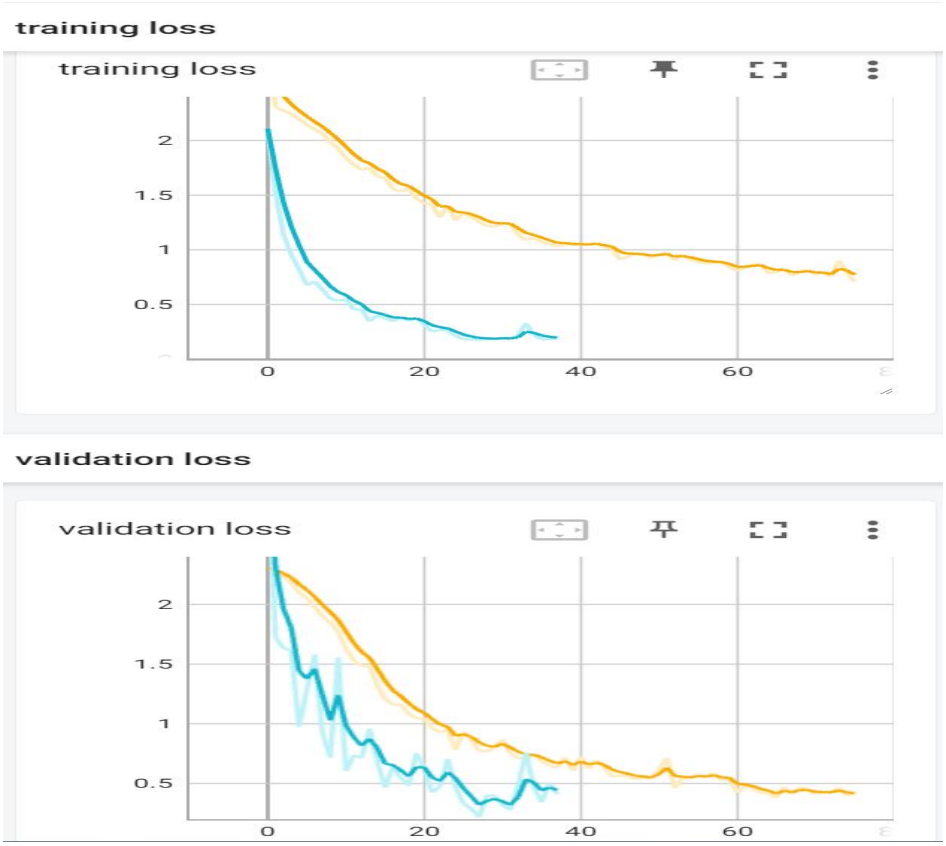
Figure[11]-ResNet Confuse Martix

Evidently, both models exhibit signs of overfitting. When employing a camera for real-time gaze direction detection, we encountered scenarios consistent with our data: the models were particularly sensitive to certain directions, accurately detecting the current focus of the subjects. For targets that moved significantly in either the horizontal or vertical direction, the models were fairly proficient in recognition. However, for gaze directions that featured smaller movements, the models failed to accurately identify and provide feedback. On another note, given the constraints of the dataset size, the final performance of these two models might represent the limit of this study. Although the performance of our custom model didn't quite match that of the pre-trained ResNet18, which is more efficient at extracting image features, the gap was not substantial. This indirectly validates the feasibility and potential for improvement of our experiment's concept.

4 FUTURE PLAN AND SPECULATIONS

4.1 Future Plan:

There remains significant room for improvement in this research. In my perspective, substantial data should be collected on actual application devices, such as AR or MR glasses, which could provide real-time, multi-directional data on eye movements. This could provide a consistent and common data collection environment that could greatly enhance the quality of our training set. In addition, a broader consideration needs to be made on the selection of the model. As seen from this research, ResNet network performs better in terms of convergence time and training cost compared to our own model[Figure 12]. In subsequent research, we might consider trying a Forward-forward neural network architecture (Hinton,G.,2022b). Of course, substantial training data is also indispensable.



[Figure 12-comparation of both network]

4.2 Speculations:

The core idea throughout this thesis is to interact with the 3x3 grid displayed on the device through the eyes. As is well known, the T9 keyboard or input method invented by Tegic Communications was popular worldwide in the late 1990s. With the advent of touchscreen mobile phones, most phones still retain this input method. In recent years, researchers have continued to further study the T9 input method (Kuang et al.,2023). On certain devices, enabling people with mobility impairments to interact with the 9-grid application on the device via eye tracking technology, as mentioned at the beginning of this thesis, could potentially serve as a valuable means of interaction and communication. Moreover, applying eye tracking technology from machine learning to everyday eye-related platforms could be an interesting trend.

5 CONCLUSION

This thesis thoroughly explain how to utilize neural networks to convolute eye images and determine labels, with the results being used to illuminate the corresponding squares in Nine-grid regions. The process encompasses each step, from conceptualizing the research, implementing experiments, data collection and its use, to the selection of neural networks, network construction, and comparison of actual results. This thesis finally proposes potential improvements for this research and forecasts for the future application scenarios of this technology.

REFERENCES

[1] Holmqvist, K., Nyström, M., & Mulvey, F. (2012, March). Eye tracker data quality: What it is and how to measure it. In Proceedings of the symposium on eye tracking research and applications (pp. 45-52).

[2]Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4511-4520).

[3] Robinson, D. A. (1964). The mechanics of human saccadic eye movement. The Journal of physiology, 174(2), 245.

[4] Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. arXiv preprint arXiv:2212.13345.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Kuang, E., Chen, R., & Fan, M. (2023). Enhancing Older Adults' Gesture Typing Experience Using the T9 Keyboard on Small Touchscreen Devices. arXiv e-prints, arXiv-2303.

[7] Hijazi, S., Kumar, R., & Rowen, C. (2015). Using convolutional neural networks for image recognition. Cadence Design Systems Inc.: San Jose, CA, USA, 9(1).

[8] Cipresso, P., Meriggi, P., Carelli, L., Solca, F., Meazzi, D., Poletti, B., ... & Silani, V. (2011, May). The combined use of Brain Computer Interface and Eye-Tracking technology for cognitive assessment in Amyotrophic Lateral Sclerosis. In 2011 5th International conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops (pp. 320-324). IEEE.

[9] Koskas, P., & Héran, F. (2013). Towards understanding ocular motility: III, IV and VI. Diagnostic and interventional imaging, 94(10), 1017-1031.

[10] Yang, Y., Li, T., Li, W., Wu, H., Fan, W., & Zhang, W. (2017). Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In Medical Image Computing and Computer Assisted Intervention−MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20 (pp. 533-540). Springer International Publishing.

[11] Alparslan, K., Alparslan, Y., & Burlick, M. (2020). Towards evaluating driver fatigue with robust deep learning models. arXiv preprint arXiv:2007.08453.

APPENDIX 1

Informed Consent Form

Study Title: Eye Gaze Analysis Through Machine Vision in Response to Content Changes in a 3x3 Grid

Introduction:

You are invited to participate in a research study. The purpose of this study is to analyze eye movements through machine vision as participants fixate on content changes within a 3x3 grid. This document provides information about the study to help you make an informed decision on whether or not to participate.

Objective:

The primary objective of this study is to analyze the variation in eye movements as participants observe changes in a 3x3 grid using machine vision technology.

Procedure:

You will enter the designated area and either wear the designated equipment or stand in front of it. You should maintain a distance of 2 to 3 cm between your eye and the equipment.

The researcher will provide detailed instructions regarding the procedure before the experiment begins to ensure the validity of the data collected.

Data Collection and Confidentiality:

The data collected during this study is solely for research purposes and will not be used for any commercial applications.

All data will be stored securely and kept confidential to ensure your privacy and safety.

Participation and Withdrawal:

Your participation in this study is entirely voluntary. You may choose not to participate or withdraw from the study at any time without any consequences.

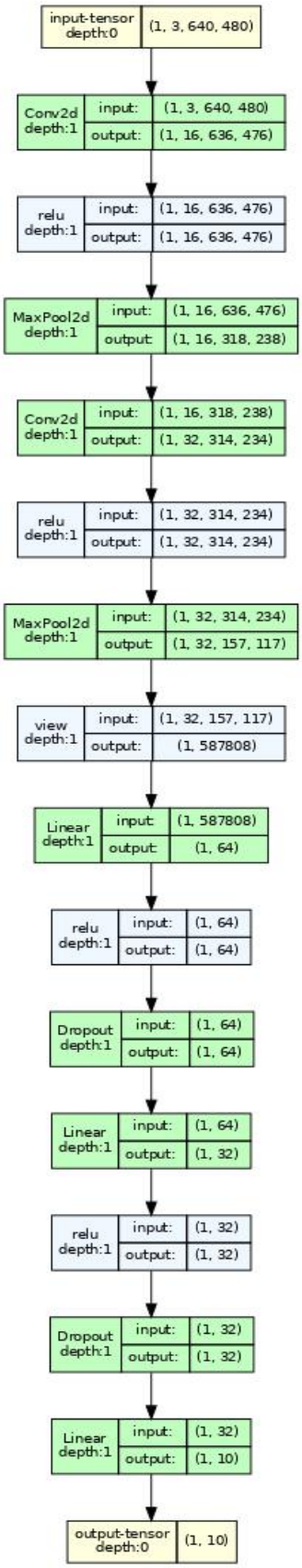
Risks and Benefits:

There are no known risks associated with participating in this study. The benefits include contributing to research in machine vision and eye movement analysis.

Consent:

I have read and understood the information provided above. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction. I voluntarily agree to participate in this study.

APPENDIX 2


input-tensor depth:0
(1, 3, 640, 480)
Conv2d depth:1
input: (1, 3, 640, 480)
output: (1, 16, 636, 476)
relu depth:1
input: (1, 16, 636, 476)
output: (1, 16, 636, 476)
MaxPool2d depth:1
input: (1, 16, 636, 476)
output: (1, 16, 318, 238)
Conv2d depth:1
input: (1, 16, 318, 238)
output: (1, 32, 314, 234)
relu depth:1
input: (1, 32, 314, 234)
output: (1, 32, 314, 234)
MaxPool2d depth:1
input: (1, 32, 314, 234)
output: (1, 32, 157, 117)
view depth:1
input: (1, 32, 157, 117)
output: (1, 587808)
Linear depth:1
input: (1, 587808)
output: (1, 64)
relu depth:1
input: (1, 64)
output: (1, 64)
Dropout depth:1
input: (1, 64)
output: (1, 64)
Linear depth:1
input: (1, 64)
output: (1, 32)
relu depth:1
input: (1, 32)
output: (1, 32)
Dropout depth:1
input: (1, 32)
output: (1, 32)
Linear depth:1
input: (1, 32)
output: (1, 10)
output-tensor depth:0
(1, 10)