HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

Machine Learning-Driven Image Noise Robustness Enhancement for Reliable Gaze Estimation

Hyungchan Yoo[1] and Tajvir Singh[#]

[1]North London Collegiate School Jeju, Republic of Korea
[#]Advisor

ABSTRACT

This research paper presents the development of a novel Human-Computer Interaction (HCI) system that enables users to control a computer mouse through eye movements by utilizing machine learning techniques. Traditional input devices, such as a mouse or keyboard, often present accessibility challenges for individuals with physical disabilities or limited mobility. To address this problem, I proposed machine learning-driven gaze estimation system for human computer interaction for quadriplegia patients. The proposed system takes both eye images as input to predict gaze vector. The system interprets real-time eye movement data to execute cursor movements which offers an alternative and inclusive method of computer interaction. To improve the accuracy of the gaze estimation system, I introduced a random gaussian noise-based denoising autoencoder. Experimental results demonstrate that this approach significantly enhanced accuracy reducing the angular error by 4.7 degrees on a public gaze estimation dataset. Additionally, I conducted real-world experiments to evaluate the model's performance in tracking the gaze of actual users. The results indicate that the proposed model achieved an accuracy of 91%, representing an enhancement of up to 15% compared to state-of-the-art gaze estimation methods.

Introduction

Human-Computer Interaction (HCI) refers to systems that facilitate interaction between users and computers which integrates a wide range of technologies and interfaces designed to optimize user experience and usability. Common examples include optical mice, keyboards, and touch screens on mobile devices. These HCI systems are typically developed for the general public, who can easily move their limbs, often involving the use of fingers for interaction. However, this reliance on physical movement poses significant challenges for individuals with limited mobility, such as those who have difficulty moving their limbs.

One such condition is quadriplegia, a disability in which a person is unable to move their limbs due to paralysis of muscles below the neck. For quadriplegic individuals, facial movement is often the only form of voluntary control they retain. Quadriplegia can result from various causes, including vehicle accidents, falls, physical violence, and sports injuries—incidents that can occur in everyday life. The prevalence of such injuries highlights the need for accessible HCI systems that enable these individuals to use computers and access the internet effectively. Developing HCI systems that are accessible to all users, including those with disabilities, presents unique challenges due to the physical limitations involved. One promising solution is gaze estimation technology, which tracks a person's eye movements to determine their gaze direction. This technology can provide quadriplegic patients with a way to interact with computers more easily by using their eye movements to control the cursor.

To address this need, I proposed a machine learning-driven gaze estimation system for HCI, specifically designed for quadriplegic patients. The proposed system uses images of both eyes as input to predict the gaze vector, allowing it to interpret real-time eye movement data and translate it into cursor movements. This approach offers an alternative and inclusive method of computer interaction. To enhance the accuracy of the gaze estimation system, I

introduced a random Gaussian noise-based denoising autoencoder, which further refines the prediction of gaze direction.

The remaining chapters of this research paper are organized as follows: Chapter 2 reviews related work to provide a deeper understanding of the proposed approach. Chapter 3 details the proposed method, including the training and testing strategies. Chapter 4 presents various experimental results and findings. Finally, Chapter 5 summarizes the paper and its contributions.

Related Work

Gaze Estimation

There have been multiple attempts to integrate gaze estimation technology into HCI systems in the past. However, creating a system with consistent performance proved challenging due to various factors, such as skin color or lighting conditions, which could interfere with the system's ability to accurately analyze the necessary gaze data.

In 2019, Park et al.proposed a method that could potentially address these challenges by effectively isolating relevant gaze information (Park et al. 2019). His approach, known as FAZE estimation, utilized a rotation matrix to estimate the direction of the user's gaze even when their head was not in a frontal view. This allowed the system to predict where the user's gaze would be if they were looking directly into the camera. However, the FAZE estimation method had a limitation: it relied on supervised training approach which demands a large-scale dataset. To train the machine learning models, a pre-determined labeled dataset is required. With a limited number of available datasets, it becomes difficult to train the AI with high accuracy, potentially affecting the system's overall performance.

To address this problem, Yu et al. proposed a way to create an unsupervised approach (Yu et al. 2020). Their model was able to calculate the rotation matrix using the previous information it received, removing the need to do training with a labelled dataset. The biggest flaw of this model was that it was too reliant on prior information. This may lead to an incorrect response of the system when the prior information is not accurate.

Sun et al. leverages this idea by applying activation map swapping techniques (Sun et al. 2021). Instead of calculating the gaze, they suggested isolating the gaze-related features by swapping consistent features. Since the features that are non-gaze related such as skin and eye colour will remain consistent throughout the usage of the system unlike the gaze, it would be easier to isolate the gaze-related features by swapping these features between two different images. Gideon et al. strengthened this idea by adding an extra camera into the system (Gideon et al. 2022). This would provide more consistent features to be swapped and hence, allow the the trained model to better isolate consistent features.

Denoise Auto-Encoder

A significant challenge in implementing gaze estimation in HCI systems is the presence of Gaussian noise in images. Gaussian noise refers to random variations in pixel values, which can affect the color or brightness of an image, often leading to a reduction in image quality. This noise complicates the process of calculating gaze direction by making it harder to identify critical features, such as the pupil, thereby significantly decreasing the system's accuracy.

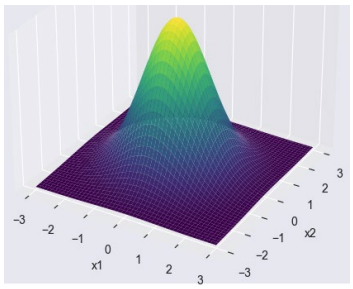


Figure 1. 3D representation of Gaussian noise

A potential solution to this problem is the Denoising Autoencoder. This system is designed to remove noise from data, enhancing the quality of the input image. The autoencoder consists of two main components: the encoder and the decoder. The encoder is responsible for filtering out noise from the image and isolating the gaze-related features. The decoder then reconstructs the image using the feature maps generated by the encoder. The reconstructed image is free from the original noise, allowing the gaze estimation system to accurately calculate the eye position without interference from noise.

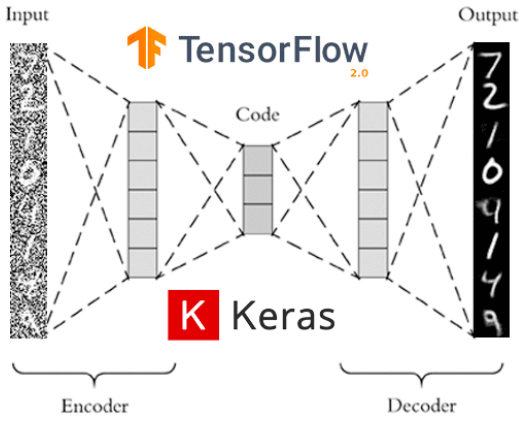


Figure 2. Explanation of denoise auto encoder (Rosebrock 2020).

Proposed Method

The proposed system is designed to be robust against random Gaussian noise in images, enabling more accurate gaze estimation. This gaze estimation system is intended for use in an HCI system that relies on a person's gaze to control a computer. It is particularly beneficial for individuals with limited limb movement, such as quadriplegic patients, who face challenges using traditional HCI systems that require finger movements.

Unsupervised Gaze Representation Learning

The system is developed with the unsupervised learning to enhance its performance. This approach allows the model to optimize its algorithm without relying on labeled datasets. Figures 3 and 4 illustrate how unsupervised learning is applied to specific components of the system.

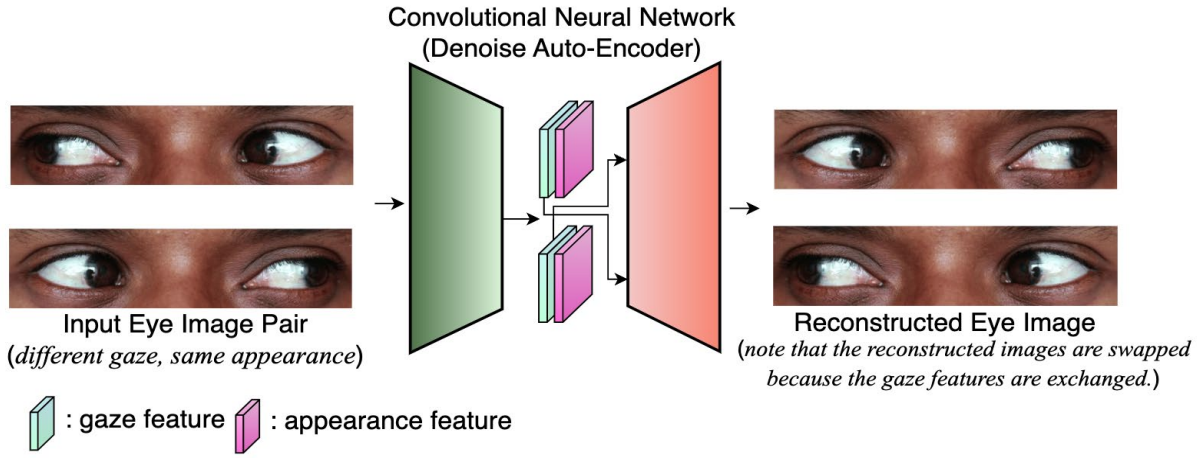


Figure 3. Architecture of the denoise auto-encoder

Figure 3 illustrates the proposed unsupervised gaze representation learning for isolating gaze-related features. During the training process of the Denoise Auto-Encoder, the system will take two different images of the user's eye looking at different places. The two images will have distinct gaze-related features but similar appearance features. The encoder will then create feature maps for gaze-related features and appearance features, and swap the gaze-related features of the images. The reconstructed images will be compared to the inputs to calculate the loss and adjust the parameters. This process is repeated until a certain level of accuracy is reached and ensures the auto-encoder can isolate the gaze-related features effectively from others.

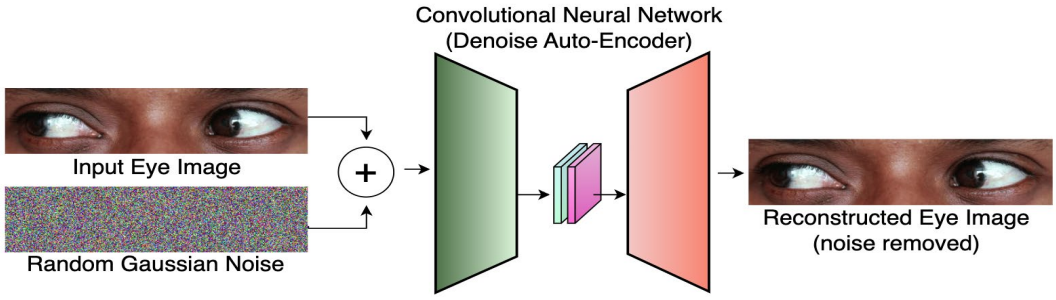


Figure 4. Architecture of the proposed denoise auto-encoder with random gaussian noise

Figure 4 displays the process of unsupervised learning to be robust against image noises. The system will again utilize the image taken from the user for training. This time, a random Gaussian noise will be inputted together with the image of the eye. The Denoise Auto-Encoder will take out the noise-related features from the produced feature maps and reassemble the image back together. Then the reconstructed image is compared with the original image of the eye to again calculate the loss and adjust the parameters accordingly. Same as before, the process is repeated to achieve a certain level of accuracy. This will allow the auto-encoder to remove noises from the input when implemented into HCI systems.

To train the proposed system, I utilized reconstruction L1 loss function often used for denoise task. The equation computes the disparity between the reconstructed image and its ground truth image as shown in Equation 1.

Equation 1. Reconstruction L1 loss function

$$L_{recon} = \frac{1}{XY}\sum_y \sum_x \left| I_{gt}(x,y) - \hat{I}(x,y) \right|$$

In Equation 1, *X* and *Y* denote the width and height of both the reconstructed and original images, respectively. *I(x, y)* represents the pixel intensity at the specified coordinates *x, y*.

Gaze Estimation

To predict the direction of gaze, the system calculates the yaw and pitch angles of the pupil in relation to the position of the head. Yaw refers to the horizontal angle of the gaze which represents how far the eye has moved left or right from its original position. Pitch represents the vertical angle of the gaze, showing how much the eye has moved up or down from its initial position. The process of calculating these angles is illustrated in Figure 5.

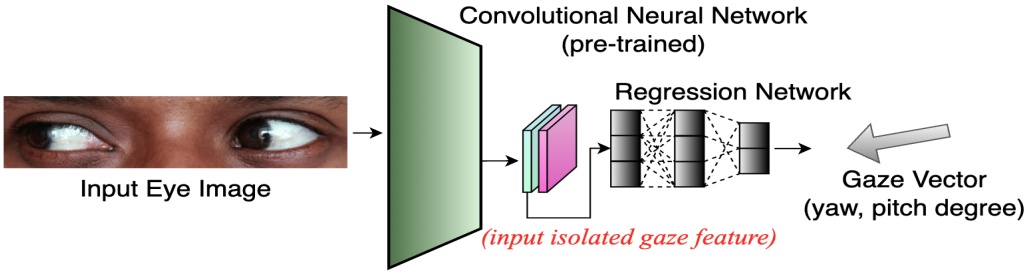


Figure 5. Architecture of the proposed gaze estimation network

Figure 5 illustrates how the pre-trained Convolutional Neural Network (CNN) calculates the yaw and the pitch angle of the eye. When the image of the eye is inputted into the network, the system will remove random Gaussian noise and isolate the gaze-related feature from other features using multiple convolution layers. Then the gaze-related feature map will be inputted into the linear regression network to calculate the yaw and the pitch angle of the gaze.

The gaze estimation network is trained with mean squared error function. The predicted yaw and pitch degree is compared to its ground truth to compute the error as explain in Equation 2.

Equation 2. Mean Squared Error Function

$$L_{gaze} = (pitch_{gt} - \widehat{pitch})^2 + (yaw_{gt} - \widehat{yaw})^2$$

Experimental Results

The dataset used for the experiment was the X-Gaze dataset created by Zhang et al. (Zhang et al. 2020). This dataset comprises 1,083,492 image samples collected from 110 participants with diverse characteristics. Among the participants, 47 were female and 63 were male, representing a range of races, including Caucasian, Middle Eastern, East Asian, South Asian, and African. Additionally, 17 participants wore contact lenses, while another 17 wore eyeglasses during image capture. By incorporating such a variety of features, the dataset enhances the system's adaptability for real-world applications in computer interaction.

Figure 6. X-Gaze dataset (Zhang et al. 2020)

The system calculated the loss using the formula below, which outputs the difference in the angle of the predicted gaze vector and the actual gaze vector in degrees. The formula uses the rearranged vector dot product formula to calculate the difference in angles between the two vectors in radians, then converts it to degrees by multiplying $\frac{180}{\pi}$.

Equation 3. Angular Error

$$\theta = \frac{180}{\pi} \times cos^{-1}\left(\frac{\overrightarrow{g_{gt}} * \overrightarrow{g_{pred}}}{\sqrt{\overrightarrow{g_{gt}}} \times \sqrt{\overrightarrow{g_{pred}}}}\right)$$

During the testing process, two different CNN were used, which were VGG-19 (Simonyan et al. 2014) and Resnet-18 (He et al. 2016). These two models were chosen as they were the best models for real-time operation. The result of the experiment is displayed in Table 1 below.

Table 1. Angular error comparison

	Angular Error (VGG-19)	**Angular Error (Resnet-18)**
FAZE (Park et l. 2019)	16.5	13.6
Gaze Redirection (Yu et al. 2020)	14.6	12.2
Cross Encoder (Sun et al. 2021)	12.3	10.7
Multi-View Cross Encoder (Gideon et al. 2022)	11.9	9.8
Proposed Method	11.3	8.9

The four previously mentioned models, along with the proposed method, were tested to compare their performance. Overall, the results demonstrate that the proposed improvements to the gaze estimation model led to a reduction in angular error. Among the four existing models, the Multi-View Cross Encoder exhibited the lowest angular error, with values of 11.9° for VGG-19 and 9.8° for ResNet-18. The proposed method improved performance for both CNN architectures, decreasing the angular error by 0.6° for VGG-19 and 0.9° for ResNet-18 when compared to the Multi-View Cross Encoder.

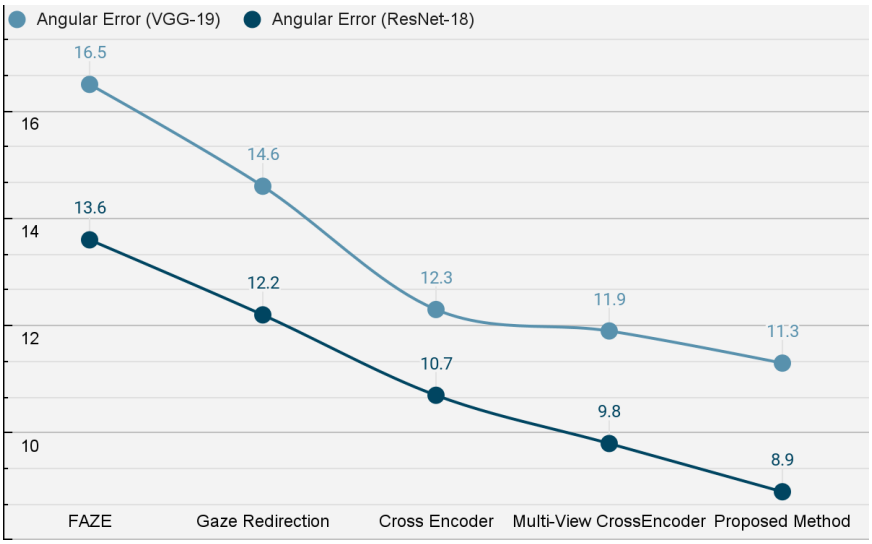


Figure 7. Angular error comparison

Figure 7 is the graphical representation of the results. It clearly shows that compared to the existing methods, the proposed method shows the greatest accuracy in gaze estimation. The gradient of both graphs are almost linear, which proves that the proposed method has improved the gaze estimation model as much as the other models and that the improvement made was meaningful.

Table 2. Ablation study for different noise generation methods

	Angular Error (VGG-19)	**Angular Error (Resnet-18)**
Proposed Method (gaussian noise)	11.3	8.9
Proposed Method (random scratch)	12.2	10.0
Proposed Method (random masking)	11.6	9.6
Proposed Method baseline)	11.9	9.8

There was also a data augmentation experiment done in order to increase the performance of the model even further. Three different geometrical patterns were added to the dataset: Gaussian noise, random scratch, and random masking. The datasets were then used for training the model. This was done to deliberately increase the level of difficulty for the system, and thus, increase the overall performance of the model. The result is displayed in Figure 9. Dataset with random scratch showed the opposite effect of increased angular error for both CNNs. Random masking did enhance the performance by 0.3° for VGG-19 and 0.2°. However, the addition of gaussian noise showed the best increase in performance by improving the angular error by 0.6° for VGG-19 and 0.9° for Resnet-18.

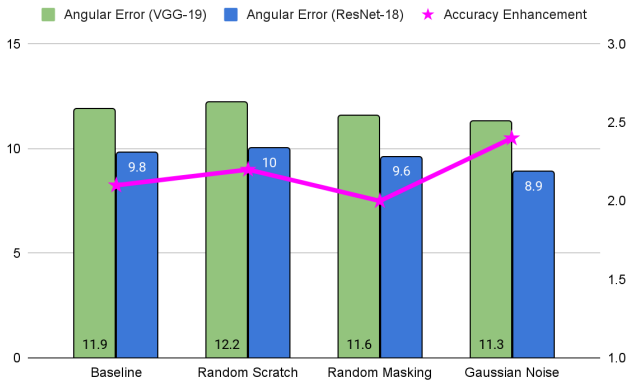
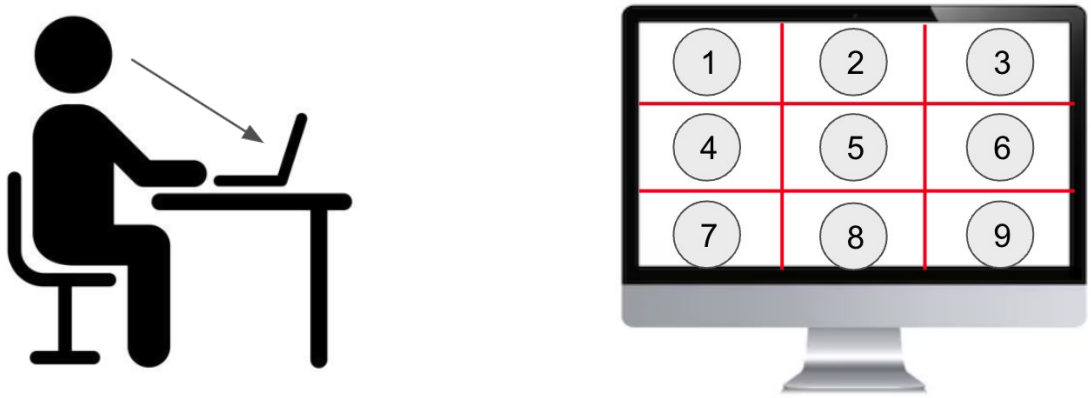


Figure 8. Ablation study for different noise generation methods

Figure 8 is the graphical representation of the result of the data augmentation experiment. It also displays how much the accuracy of the system was enhanced through each method compared to the base line. The graph shows clearly that the experiment that used Gaussian Noise had the most significant enhancement in accuracy of almost 2.5% from the baseline.



(a) (b)

Figure 9. User study setting
(a): individual looking at the monitor screen and (b): monitor screen with 9 splitted sections

After training the system, there was an experiment done on actual people that might use the system to evaluate the performance when it is applied to the computer. All of the four existing methods and the proposed method was used. For the procedure of the experiment, the test subjects were only allowed to use their eyes. Then the screen would randomly display the nine dots in random order and the test subject is instructed to look at the dots one by one as they appear. Everytime the test subject looks at a dot, the system will calculate the yaw and pitch of the pupil, then project it on the screen. If the projection is within the region of the section as shown below, the attempt of estimating the gaze is considered successful. For example, if the user was instructed to look at dot number 5, and the projected result is within the fifth square, the attemp is successful. This was repeated 3 times for 5 different test subjects.

Table 3 shows the result of the accuracy for each of the methods. The calculation of accuracy was done by dividing number of successful attempts by the number of total attempts. Again, the experiment was done using two of

the CNNs used before. For both CNN, the proposed method showed the greatest level of accuracy out of all the models with 80.74% for VGG-19 and 91.11% for Resnet-18.

Table 3. User study evaluation

	Angular Error (VGG-19)	**Angular Error (Resnet-18)**
FAZE	0.7037	0.7629
Gaze Redirection	0.7333	0.7851
Cross Encoder	0.7851	0.8370
Multi-View Cross Encoder	0.7925	0.8814
Proposed Method	0.8074	0.9111

Figure 10 is the graphical representation of the result above. This graph also shows the performance gap between each model. As shown, there as been around 4% enhancement in accuracy between the Multi-View Cross Encoder model and the proposed model when ResNet-18 was used, and around 2% enhancement in accuracy for VGG-19. These enhancements are comparable to the enhancements made before and shows that the proposed method can be used in real-life situations with high accuracy.

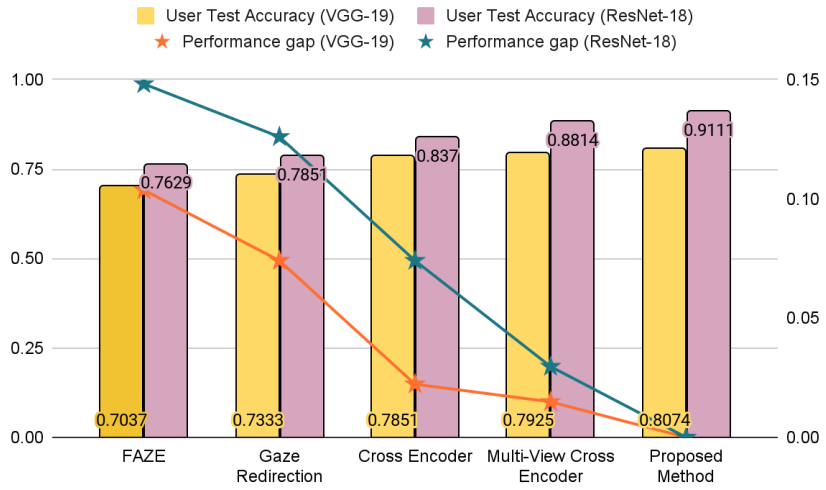


Figure 10. User study evaluation

Conclusion

In this research, I proposed a gaze estimation-driven human-computer interaction (HCI) system for individuals with quadriplegia, who face difficulties using computers due to the reliance on limb movement in most HCI systems. To enhance the accuracy of gaze estimation, I incorporated random Gaussian noise into the dataset, allowing the system to become more robust against such noise. This addition significantly improved the proposed method's accuracy. The experimental results demonstrated that this approach could be effectively utilized as an HCI solution for quadriplegic patients, achieving a minimum angular error of 9.8°, marking a meaningful advancement in gaze estimation. Additionally, I conducted real-world experiments to assess the model's performance in tracking the eye movements of

actual users. The results indicated that the proposed model achieved an impressive accuracy of 91%, with an enhancement of up to 15% compared to state-of-the-art gaze estimation methods. In the future, I plan to develop a blink detection system to further enhance the features and usability of the HCI system.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Gideon, J., Su, S., & Stent, S. (2022). Unsupervised multi-view gaze representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5001-5009). https://doi.org/10.1109/CVPRW56347.2022.00548

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). https://doi.org/10.48550/arXiv.1512.03385

Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O., & Kautz, J. (2019). Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9368-9377). https://doi.org/10.48550/arXiv.1905.01941

Rosebrock, A. (2020, Feb 24). “Denoising autoencoders with Keras, TensorFlow, and Deep Learning”: Py Image Search
https://pyimagesearch.com/2020/02/24/denoising-autoencoders-with-keras-tensorflow-and-deep-learning/

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Sun, Y., Zeng, J., Shan, S., & Chen, X. (2021). Cross-encoder for unsupervised gaze representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3702-3711). https://doi.org/10.1109/ICCV48922.2021.00368

Yu, Y., & Odobez, J. M. (2020). Unsupervised representation learning for gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7314-7324). https://doi.org/10.48550/arXiv.1911.06939

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 (pp. 365-381). Springer International Publishing.