HIGH SCHOOL EDITION
Journal of Student Research

Implementation of Eye Movement-based Digital Healthcare Communication Board: Accurate Gaze Estimation Using Triplet Loss Function

Christopher Shin[1], Jihan Lee[2] and Youngjee Kim[#]

[1]Korea International School, Republic of Korea
[2]Korea International School Pangyo Campus, Republic of Korea
[#]Advisor

ABSTRACT

Effective communication is important in healthcare, especially for quadriplegia patients who often face barriers in expressing their needs. Traditionally, these patients rely on healthcare communication boards to interact with doctors and nurses. However, this method is often slow, cumbersome, and requires the assistance of another person which makes it an inconvenient solution. Recent advancements in gaze estimation technology, which predicts the direction of an individual's eye movements, can provide a promising alternative. This research examines the potential of gaze estimation to develop a digital healthcare communication board driven by eye movements. Such a system would empower quadriplegia patients to communicate independently which enhances both the speed and efficiency of interactions. The proposed system processes eye images to predict a gaze vector which represents the direction in which an individual is currently looking. To improve the system's accuracy, we introduce a conjugate ability-based loss function. Additionally, the proposed approach was applied to a digital healthcare system to demonstrate its feasibility and effectiveness in real-world scenarios. The system achieved an angular error of 8.7 on a public gaze estimation dataset which surpasses previous state-of-the-art methods.

Introduction

Quadriplegia refers to a condition characterized by the paralysis of all four limbs—both arms and legs—typically resulting from a spinal cord injury or a neurological disorder. Individuals with quadriplegia may also experience varying degrees of paralysis in their torso and trunk, depending on the severity and location of the spinal injury. This condition impairs motor function and sensation below the level of injury which affects the ability to perform everyday tasks and communicate.

Quadriplegia patients often lose the ability to move or control their limbs, which severely limits their ability to speak or use traditional methods of communication, such as writing or typing. As a result, they rely on communication boards to interact with doctors, nurses, and others. A communication board is a tool that typically displays letters, words, or symbols that the patient can point to, either with their eyes or with the help of a caregiver, to express their needs, ask questions, or convey information. For many quadriplegia patients, this is one of the few ways they can actively participate in their care and communicate with others. However, these traditional communication boards, though essential for quadriplegia patients, have significant limitations. They require an assistant to point to each item one by one which makes the process time-consuming.

To address the limitations of traditional communication boards, we propose an eye movement-based digital healthcare communication board using gaze estimation. The proposed system consists of two modules: gaze representation learning and gaze estimation. In the first module, we introduce a triplet loss function that leverages conjugate gaze ability to enhance the representation of gaze-related features. The second module, gaze estimation, uses the pre-trained network to accurately predict the gaze vector, which includes both yaw and pitch angles. We implemented the proposed network in a laptop environment to develop the digital healthcare board.

The remainder of this paper is structured as follows: Chapter 2 provides a conceptual overview of Healthcare Communication Boards and gaze estimation to establish a foundation for understanding this research. Chapter 3 details the proposed approach, including the training strategy. Chapter 4 demonstrates the effectiveness of our approach through extensive experimental results. Finally, Chapter 5 summarizes the paper.

Background Knowledge

Healthcare Communication Board

Quadriplegia, also known as tetraplegia, is a type of spinal cord injury that results in paralysis of all four limbs and the body below the neck (Cleveland Clinic, 2022). The human spine consists of 33 vertebrae and 31 spinal nerves, divided into five sections: cervical, thoracic, lumbar, sacral, and coccygeal. Quadriplegia occurs when the cervical vertebrae, which include seven segments labeled C1-C7, are injured. These vertebrae start at the base of the skull and extend just above the chest. The specific location of the spinal injury determines which parts of the body are paralyzed. According to the World Health Organization (WHO), approximately 15.4 million people worldwide live with spinal cord injuries (SCI), with about 60% of cases resulting in quadriplegia.

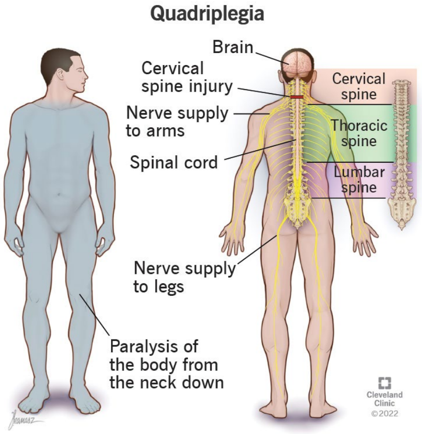
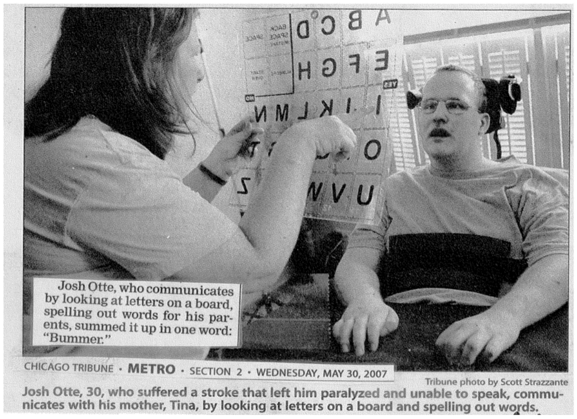


Figure 1. Different sections of the spine (Cleveland Clinic 2024).

Understanding the causes of quadriplegia is important. Unlike many other conditions, quadriplegia is typically not inherited. Most cases of quadriplegia are acquired through various incidents and accidents that occur in life events. For instance, falls and road traffic accidents are the leading cause, followed by violence from self-harm or attempted suicide. (WHO, 2023)



Josh Otte, who communicates by looking at letters on a board, spelling out words for his parents, summed it up in one word: "Bummer."

CHICAGO TRIBUNE · METRO · SECTION 2 · WEDNESDAY, MAY 30, 2007

Tribune photo by Scott Strazzante

Josh Otte, 30, who suffered a stroke that left him paralyzed and unable to speak, communicates with his mother, Tina, by looking at letters on a board and spelling out words.

Figure 2. Quadriplegia patient's parents going through letters one by one for the patient (Autodesk 2024).

To address the communication challenges faced by quadriplegics, significant efforts have been made to find effective methods for enabling them to engage in conversations. The most widely used solution is the healthcare communication board, which displays alphabet letters and short words or phrases (Figure 2). A nurse or caregiver points to the items on the board, and the patient responds with eye movements. Despite the severe physical limitations quadriplegics face, they retain control over their eye movements because the muscles responsible for eye movement are connected directly to the brain via oculomotor nerves, unaffected by spinal cord injuries. However, communication boards have significant drawbacks. They require a nurse or caregiver to hold the board and painstakingly go through the letters one by one, making the process time-consuming and slow.

Gaze Estimation

Gaze estimation is a technology used to determine where a person is looking. The output of gaze estimation is typically a gaze vector, which represents the direction of the gaze in a three-dimensional space. This vector is often decomposed into two angles: yaw and pitch.

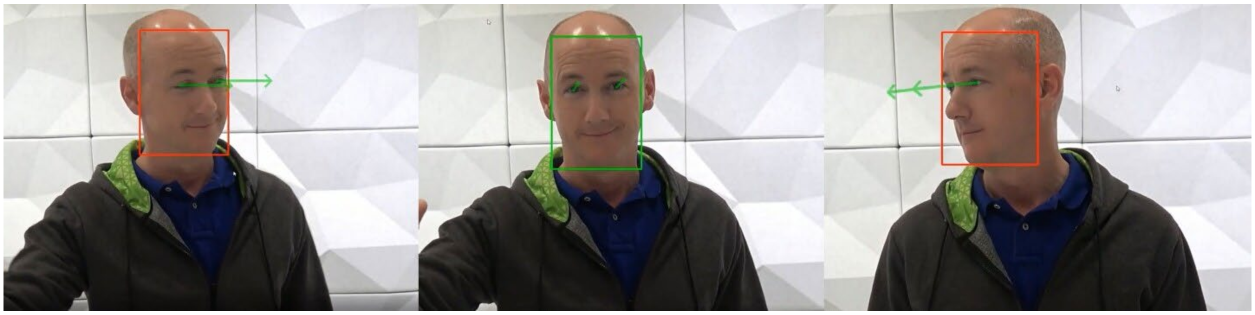


Figure 3. Gaze estimation (NVIDIA 2024)

Yaw refers to the horizontal rotation of the gaze direction relative to a forward reference direction, measured in degrees where 0 indicates looking straight ahead; positive values indicate a gaze to the right, and negative values indicate a gaze to the left (e.g., +30 degrees for 30 degrees to the right, -15 degrees for 15 degrees to the left). Pitch refers to the vertical rotation of the gaze direction relative to the horizontal plane, also measured in degrees where 0 indicates looking straight ahead; positive values indicate a gaze upwards, and negative values indicate a gaze downwards (e.g., +20 degrees for 20 degrees up, -10 degrees for 10 degrees down).

Gaze estimation is widely used in various fields, such as human-computer Interaction, virtual reality or augmented reality. In assistive technologies, it can help individuals with disabilities interact with devices using eye movements. Marketing and advertising benefit from insights into consumer behavior through gaze pattern analysis, while psychological and cognitive studies use it to understand cognitive processes and social interactions.

In Particular, gaze estimation can significantly enhance the functionality of communication boards by enabling users to select items or letters using their eye movements. In this study, we utilize gaze estimation to develop a digital healthcare communication board. Detailed information about the proposed system is further explained in Chapter 3.

Proposed Gaze Estimation Network

In this chapter, the proposed eye movement-based digital healthcare communication board system is described in detail. The system consists of three main modules: gaze representation learning, gaze estimation, and post-processing. In the gaze representation learning module, a convolutional neural network is trained using a triplet loss function to identify and enhance gaze-related features. The trained network is then utilized in the gaze estimation module to accurately predict gaze vectors from input eye images. Finally, the post-processing module determines the specific area that the individual is gazing at, based on the predicted gaze vectors.

Gaze Representation Learning

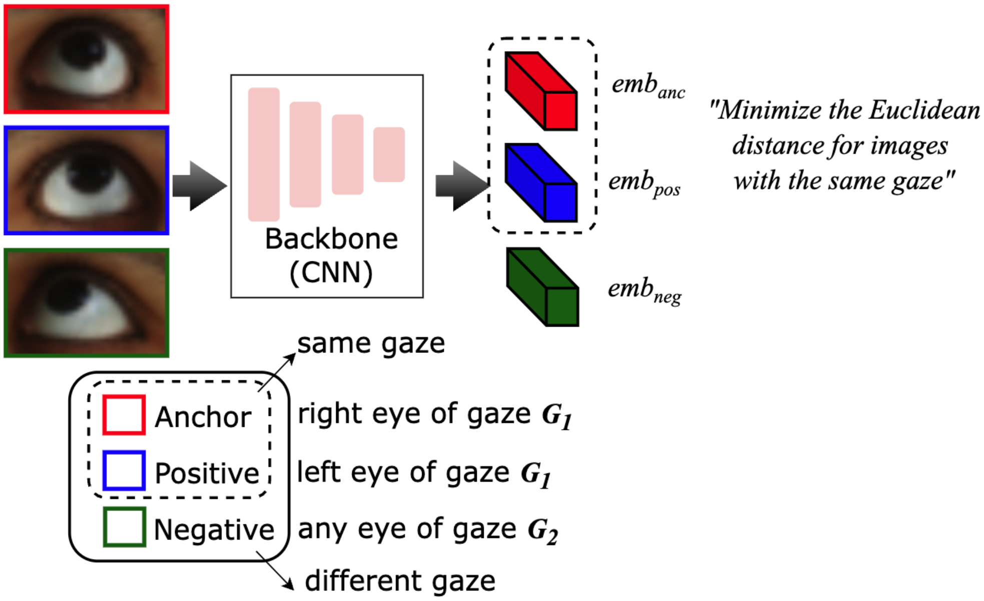


Figure 4. Architecture of the proposed gaze representation learning

Figure 4 illustrates the architecture of the proposed gaze representation learning system. The goal of this process is to train a convolutional neural network (CNN) to extract consistent gaze-related features, which are then used to further train the gaze estimation network. The CNN processes three pairs of eye images, referred to as a triplet pair. Each triplet consists of three eye images: the left and right eyes looking in the same direction at a specific time, and a third image (either left or right) looking in a different direction. In this research, the two eye images that are looking in the same gaze direction are referred to as the Anchor and Positive samples, while the eye image looking in a different direction is referred to as the Negative sample. The CNN takes this triplet pair as input and generates corresponding feature maps, or embedding vectors, which mathematically represent the gaze-related features in the input images.

The hypothesis of the proposed approach is that the embedding vectors of the Anchor and Positive samples are mathematically similar, as they have the same gaze direction, while the embedding vectors of the Anchor and Negative samples are dissimilar, due to their different gaze directions. To implement this, we first measure the discrepancy between each pair of embedding vectors using Equation 1.

Equation 1: Distance function

$$D_{i,j} = \sqrt{(emb_i^1 - emb_j^1)^2 + (emb_i^2 - emb_j^2)^2 + \cdots + (emb_i^F - emb_j^F)^2}$$

Here, emb^k denotes the k-th element of the embedding vector, ***F*** represents the dimension of the embedding vector, and finally $D_{i,j}$ explains the computed distance between two vectors i and j. Equation 1 measures the Euclidean distance between two ***F***-dimensional embedding vectors. This discrepancy is referred to as the "distance" in this study, and we compute two distances: one between the Anchor and Positive pair and another between the Anchor and Negative pair. These distances are then used to formulate a hinge loss function, as explained in Equation 2.

Equation 2: Hinge loss function

$$J_{hinge} = \max\,(0, D_{anc,pos} - D_{anc,neg} + \mathrm{M})$$

Equation 2 defines the loss function used to train the proposed gaze representation learning model. To minimize the loss value, the distance between the Anchor and Positive embedding vectors should be minimized, while the distance between the Anchor and Negative embedding vectors should be maximized. The function measures the difference between these two distances and returns 0 if the difference exceeds the variable *M*. Here, *M* is the marginal threshold parameter that controls the strength of the training strategy. In this study, *M* is set to 1.

Gaze Estimation

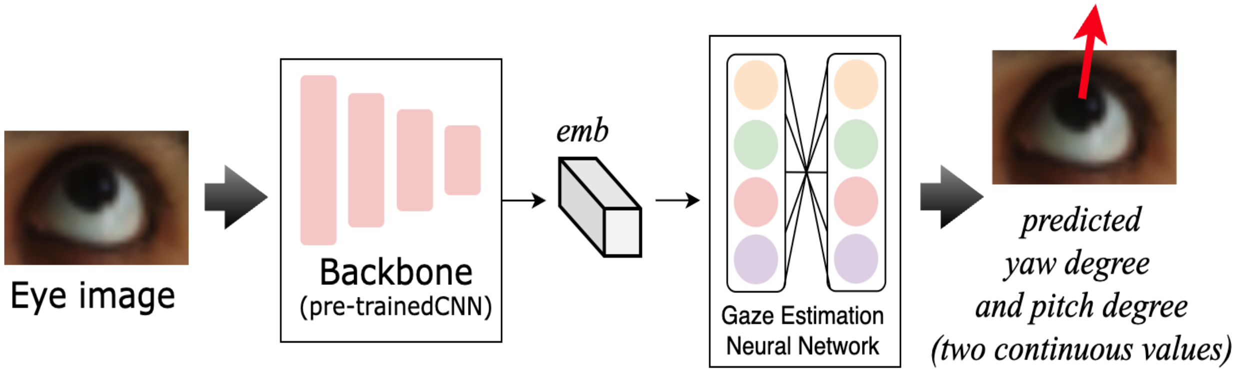


Figure 5. Architecture of the proposed gaze estimation network

Figure 5 demonstrates the process of gaze estimation where the input data is processed through the backbone and the gaze estimation neural network, ultimately producing the final output. The input for the whole procedure of gaze estimation is a single eye image of the patient while the output is a single value, the angular error, measured in degrees. This represents the angle between the predicted and actual gaze directions. Through transfer learning, a portion of the pre-trained CNN, also known as the backbone, from the gaze representation learning will be transferred in order to produce accurate and convenient outputs. Utilizing the pre-trained CNN enables them to extract gaze-related rich features. After acquiring the embedding vector from the pre-trained CNN, the embedding vector should be inputted into the gaze estimation neural network which uses the loss function to calculate the angle between the predicted yaw and pitch vector and its corresponding ground truth vector. The result will become the output represented as a degree where the yaw degree is the vertical movement of the eyes while the pitch degree is the horizontal movement of the eyes. The effectiveness of the proposed approach is further explained in Chapter 4.

To compare the predicted yaw and pitch vector with its corresponding ground truth vector, the angular error function is used for gaze estimation neural networks as shown in Equation 3.

Equation 3: Angular error function

$$J_{gaze} = \frac{180}{\pi} \times cos^{-1}\left(\frac{yaw \times \widehat{yaw} + pitch \times \widehat{pitch}}{\sqrt{yaw^2 + pitch^2} \times \sqrt{\widehat{yaw}^2 + \widehat{pitch}^2}}\right)$$

In the previously mentioned Equation 3, yaw and pitch denote the ground truth vector for both yaw and pitch, while $\widehat{yaw}$ and $\widehat{pitch}$ designate the predicted yaw and pitch vector. J_{gaze} indicates the output value that is represented as a degree of the angle between the predicted and the actual gaze directions. To obtain the most accurate result, the difference between the original yaw and pitch vector and the predicted yaw and pitch vector should be close to zero, which means higher accuracy. If the output value in degrees is lower, it shows us that the predicted yaw and pitch vector is very accurate, pointing essentially the same direction which means that the loss is low.

Experimental Results

Gaze Capture Dataset

Gaze Capture Dataset (Krafka et al. 2016) is a large-scale dataset containing data from over 1450 people consisting of almost 2.5 million frames. This dataset contains variations of each category. For instance, there are various races, genders, and objects on their face like glasses. In addition, each photo has distinct lighting, background, and even their face position in the image.

Figure 6. Image examples from the Gaze Capture Dataset (Krafka et al. 2016).

Evaluation Protocol

To statistically validate the experiment, we conducted 5-fold cross-validation. In this process, the dataset is divided into five equal parts. Each part is used as a test set while the remaining four parts are used for training the model. This process is repeated five times, with each part used as the validation set once. The results from each iteration are averaged to provide a more reliable estimate of the model's performance.

To access the accuracy of the proposed network, we measured the angular error between the predicted gaze vector and its ground truth vector.

Equation 4: Angular Error

$$Error = \frac{180}{\pi} \times cos^{-1}(\frac{yaw \times \widehat{yaw} + pitch \times \widehat{pitch}}{\sqrt{yaw^2 + pitch^2} \times \sqrt{\widehat{yaw}^2 + \widehat{pitch}^2}})$$

Equation 4 computes the angle in degrees between two vectors, each composed of yaw and pitch. If the two vectors have identical directions, the angle is zero degrees; if the vectors are opposite, the angle is 180 degrees.

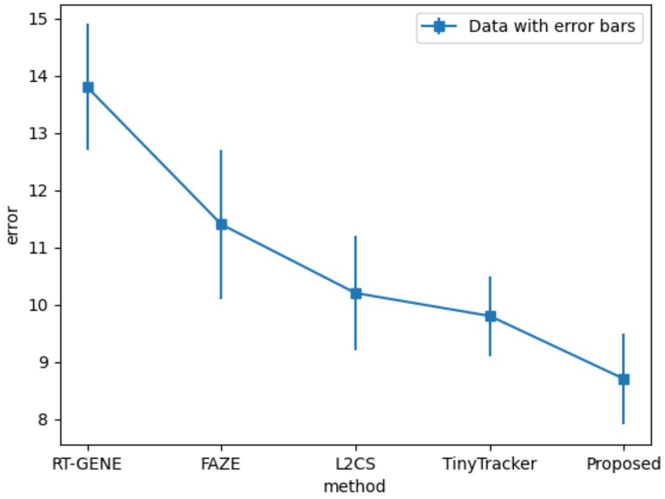
Evaluation Result

Table 1 shows a performance comparison between the proposed approach and previously developed gaze estimation methods. To ensure a fair comparison, identical experimental setups were used for both training and testing.

The proposed approach achieved state-of-the-art performance with an angular error of 8.7 degrees, surpassing all four comparison methods.

Table 1. Angular error comparison with previous gaze estimation methods

Method	Angular Error
RT-GENE (Fischer et al. 2018)	13.8
FAZE (Park et al. 2019)	11.4
L2CS (Abdelrahman et al. 2023)	10.2
TinyTracker (Bonazzi et al. 2023)	9.8
Proposed approach	8.7



Figured 7. Angular error with error bars

Figure 7 illustrates the results of the 5-fold cross-validation for all five methods. As depicted, the results are statistically reliable, with a very small standard deviation indicating consistent performance. Additionally, to examine the effectiveness of the triplet loss function, we conducted an ablation study. We trained two models using different vector similarity metrics: distance-based and angular-based. For the distance-based approach, we measured the similarity between two vectors using Equation 1, while for the angular-based approach, we used Equation 5 to measure similarity. Additionally, we varied the threshold angle degree for selecting triplet pairs to evaluate how accuracy changes for different triple pair sets. When the threshold angle is set to 40 degrees, it means that the angle difference between the anchor-positive pair and the anchor-negative pair is at least 40 degrees.

Equation 5: Angle-based distance function

$$D_{i,j} = 1 - \frac{emb_i * emb_j}{|emb_i|_1 \times |emb_j|_1}$$

Here, the * denotes the dot product, while |.| represents the L1 norm. The result of the equation is 0 if the two vectors have identical directions, and 2 if the two vectors have opposite directions.

Table 2. Ablation study result

	T=40°	**T=20°**	**T=10°**	**T=5°**	**Baseline**
Proposed approach (distance based)	8.7	9.6	11.6	13.4	**13.6**
Proposed approach (angle based)	8.9	10.8	12.9	13.5	13.6

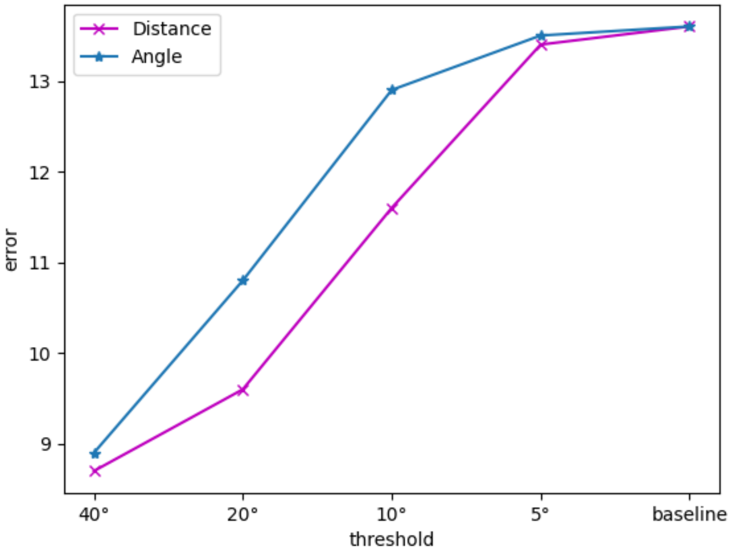


Figure 8. Ablation study result (angular error with different threshold)

Table 2 and Figure 8 summarize the results of the ablation study. Notably, higher threshold values consistently yield more accurate results for both ablation models. This indicates that the quality of the triplet pairs is directly correlated with the model's accuracy. For the metric experiment, the distance-based approach slightly outperforms the angular-based approach across all training scenarios. Surprisingly, the baseline, training the gaze estimation network from scratch without the proposed transfer method, yielded significantly poor results. This clearly demonstrates the effectiveness of the proposed approach.

Conclusion

In this paper, we propose a gaze estimation-based digital healthcare communication board. To enhance the accuracy of gaze estimation, we introduced a triplet loss function which achieved state-of-the-art accuracy on a public gaze estimation dataset. We conducted an ablation study to further examine the effectiveness of the proposed triplet loss function by comparing the angular errors between distance-based and angle-based models. The experimental results showed that the distance-based approach yields the best results when the threshold angle is greater than 40 degrees for the triplet pair setup. In the future, we plan to implement the proposed system in real-world scenarios, such as hospitals and nursing facilities.

Acknowledgments

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

References

Abdelrahman, A. A., Hempel, T., Khalifa, A., Al-Hamadi, A., & Dinges, L. (2023, October). L2cs-net: Fine-grained gaze estimation in unconstrained environments. In 2023 8th International Conference on Frontiers of Signal Processing (ICFSP) (pp. 98-102). IEEE.

Autodesk. (2024, Aug 27). "*Communication Board for Individuals With Disabilities*": Autodesk. https://www.instructables.com/Communication-Board-for-Individuals-with-Disabilit/

Bonazzi, P., Rüegg, T., Bian, S., Li, Y., & Magno, M. (2023, October). Tinytracker: Ultra-fast and ultra-low-power edge vision in-sensor for gaze estimation. In 2023 IEEE SENSORS (pp. 1-4). IEEE.

Cleveland Clinic. (2024, Aug 27). "*Quadriplegia*": Cleveland Clinic. https://my.clevelandclinic.org/health/symptoms/23974-quadriplegia-tetraplegia

Fischer, T., Chang, H. J., & Demiris, Y. (2018). Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European conference on computer vision (ECCV) (pp. 334-352).

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

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2176-2184).

NVIDIA. (2024, Aug 27). "*Gaze Estimation*": NVIDIA. https://docs.nvidia.com/tao/tao-toolkit/text/model_zoo/cv_models/gazenet.html

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).