Leveraging Computer Vision to Establish a Correlation between Eye-Gaze Estimation and Saccades

Aarav Sharma[1] and Isabel Hyo Jung Song[#]

[1]Archbishop Mitty High School
[#]Advisor

ABSTRACT

The purpose of this project was to develop a model that can detect pupils on a subject's face and draw the graph of both pupils to determine if a subject has saccades. Current methods cannot accurately detect contours when the face is presented at an angle. To resolve this issue, subjects recorded a video while performing the Head Impulse Test (HIT). Next, a computer vision library, Opencv, extracted frames from the recorded video and detected the facial key point for the nose. The Multi-task Cascaded Convolutional Neural Network (MTCNN) extracted the face from the frame and generated contours for the eyes using a segmentation library. The largest contours on the mask were divided into two parts and their extreme bounding box points were identified using dilation, erosion, and blur. Our model rendered a video of both contours applied to the patient performing the HIT test. Also, the model generated two graphs, comparing each eye's gaze with the pose estimation. There were two main outputs: the graph may resemble a y = -x line if a patient does not have saccades or the graph may be more distorted if the patient has saccades due to a sudden shift in eye-gaze movement. This study demonstrates a precise method of eye-gaze estimation and the detection of saccades. Further experiments will help to validate the notion that saccades are connected to neurological disorders. Future experiments may include more data during the training process and to quantitatively determine the accuracy of our model.

Engineering Goals and Purpose

Saccades are eye movements that quickly shift the eye's focus between two fixed points. They are used any time that your gaze moves from one point of gaze fixation to another. For example, if you are reading a book when you move from word to word or transition from the end of a line to the start of the next one. Normal head movements have very high accelerations— 4,000°/s/s and above, and so the eye movement response must have a very short latency and be accurate. The response is very fast: about 8 ms from the onset of the head movement stimulus to the onset of the eye movement response. Healthy brains and eyes can normally saccade to a new target in 1/10th of a second or less. However, brain injuries and damaged neural pathways can lead to irregular saccadic eye movements. In a way, saccades can help us to identify what brain areas are injured and allow doctors to treat that part of the brain. Some common symptoms of saccades are memory loss, forgetfulness, anxiety, agitation, oscillopsia, canal paresis, and mood changes. Dizziness and vertigo are the most common symptoms to bring a patient to a neurologist. Vertigo is defined as a sensation of self-motion or motion of the external environment. Current methods to detect abnormal saccades include the head impulse test(HIT) and the video head impulse test(vHIT). The head impulse test is a critical component of the bedside assessment of vestibular function. The video head impulse test(vHIT) allows examiners to objectively assess the vestibulo-ocular reflex during head impulses in the plane of each semicircular canal. The purpose of the VOR is to maintain a steady vision of the head movement. In the standard vHIT, the patient is required to stare at a fixed target during the head impulse. Current methods to detect abnormal saccades such as RT-GENE and iTracker may be inaccurate, expensive, or both. This is the main reason why optometrists may be inclined toward a computer algorithm that can quickly and accurately detect abnormal saccades. The main purpose of this project is to develop an easy-to-implement algorithm that is accurate and can quickly detect abnormal saccades, allowing optometrists to treat the specific portion of the brain and potentially prevent neurological disease. The hypothesis for this project is that the new model as designed will outperform older algorithms and people in its ability to accurately diagnose saccades even when a portion of the face is not visible. The first demonstration presented will be on a series of recorded views of saccadic and non-saccadic patients from YouTube videos. The model will use the MTCNN algorithm. The MTCNN framework is a solution for both face detection and face alignment, allowing optometrists to accurately detect abnormal saccades.

Background Research

Saccades are eye movements that quickly shift the eye's focus between two fixed points. They are used any time that your gaze moves from one point of gaze fixation to another. For example, if you are reading a book when you move from word to word or transition from the end of a line to the start of the next one. Normal head movements have very high accelerations— 4,000°/s/s and above, and so the eye movement response must have a very short latency and be accurate. The response is very fast: about 8 ms from the onset of the head movement stimulus to the onset of the eye movement response. Healthy brains and eyes can normally saccade to a new target in 1/10th of a second or less. However, brain injuries and damaged neural pathways can lead to irregular saccadic eye movements. In a way, saccades can help us to identify what brain areas are injured and allow doctors to treat that part of the brain. For example, during rightward head rotations, patients who had undergone right vestibular neurectomy would make leftward saccades; during leftward head rotations, patients who had undergone left vestibular neurectomy would make rightward saccades. Our brain uses saccadic eye movements to create a constantly updated, unconscious map of our body in relation to its environment. Visual stimuli from our eyes will produce neural activity in a part of our brain called the superior colliculus, which allows us to reach for objects outside our field of vision, avoid obstacles, and balance as we move. There are several types of saccades such as volitional, predictive, antisaccades, reflexive saccades, express saccades, spontaneous saccades, and quick phases of nystagmus, each of which has its own unique function. For example, volitional saccades assist with memory whereas reflexive saccades assist with visual movement of the eyes. All of these types of saccades can detect which brain area is damaged and can act as symptoms that show neurodegenerative diseases including Alzheimer's disease, Lewy Body Dementia, Parkinson's Disease, and Huntington's Disease.

Some common symptoms of saccades are memory loss, forgetfulness, anxiety, agitation, oscillopsia, canal paresis, and mood changes. Dizziness and vertigo are the most common symptoms to bring a patient to a neurologist. Vertigo is defined as a sensation of self-motion or motion of the external environment. Patients presenting with vertigo and dizziness in the emergency department typically fall into one of the following three categories: acute severe dizziness, recurrent attacks of dizziness, or recurrent positional dizziness. A patient who presents with sudden onset severe dizziness - in the absence of prior similar episodes - has the "acute severe dizziness". Nystagmus is a term used to describe alternating slow and fast movements of the eyes. Nystagmus in vestibular neuritis is spontaneous for at least the first several hours of symptoms.

Current methods to detect abnormal saccades include the head impulse test(HIT) and the video head impulse test(vHIT). The head impulse test is a critical component of bedside assessment of vestibular function. The technique leverages a high acceleration, rapid, low amplitude head rotation to assess the integrity of the vestibulo-ocular reflex(VOR), a component that triggers eye movements in response to stimulation. In specialty practice, the horizontal HIT, also known as the h-HIT) is now widely used to assist in the clinical diagnosis of peripheral vestibular disorders. The video head impulse test(vHIT) allows examiners to objectively assess the vestibulo-ocular reflex during head impulses in the plane of each semicircular canal. The purpose of the VOR is to maintain the steady vision of the head movement. In the standard vHIT, the patient is required to stare at a fixed target during the head impulse. If their VOR is not adequate, the patient must make a corrective saccade to return fixation to the target. In both of these tests, doctors might use a special device called Phantom v2511, and while it has a higher frame rate, it is very expensive.

Current methods to detect abnormal saccades may be inaccurate, expensive, or both. This is the main reason why optometrists may be inclined toward a computer algorithm that can quickly and accurately detect abnormal saccades. Essentially, detecting saccades is a two-step process. The first step is to recognize a patient's face properly. In order to do that, we need to extract the 68 different facial landmarks.

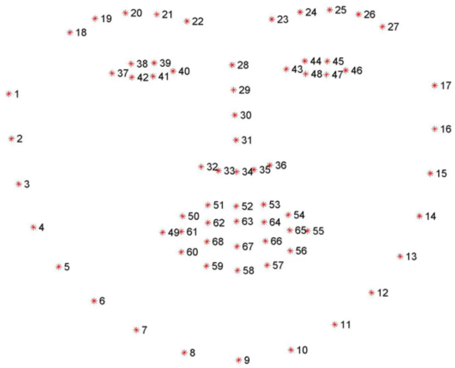


Figure 1. Key facial Landmarks

Based on this figure, we can find some important facial landmarks such as the nose tip (#34), eye pupils (#42 and #47), lips (#50-61), etc. The second step is to extract the eyes from the facial landmarks and analyze the gaze movement, also known as gaze behavior. Gaze behavior is an important non-verbal cue in social signal processing and human-computer interaction. Some notable eye gaze estimation algorithms are Real-Time Gaze Estimation in Natural Environments (RT-GENE) and iTracker. RT-GENE is a robust gaze estimation algorithm that can easily detect the eyes even in the natural environment but use expensive goggles. iTracker is a pre-trained convolutional neural network, which is used for tracking the eye position as a function of time. Despite its advantages, the iTracker algorithm with and without glasses had a horrible accuracy with a medium-light brightness set to 26 Lux.

Design, Criteria, Testing, and Evaluation

Criteria: The model must be inputted with recorded videos and can run relatively fast on a laptop. If these criteria are met, the experimenter will test it in the following manner:

1. A pre-trained model for face detection and key points called MTCNN was used. It is a combination of three different cascaded neural networks.

A. According to the authors of the MTCNN model, they "exploit a fully convolutional network, called Proposal Network (P-Net), to obtain the candidate windows and their bounding box regression vectors in a similar manner. Then [they] use the estimated bounding box regression vectors to calibrate the candidates. After that, [they] employ non-maximum suppression (NMS) to merge highly overlapped candidates."

B. After Step 1, all candidates are passed to the Refine Network (R-Net), which rejects many false candidates and performs calibration with bounding box regression.

C. Lastly, the MTCNN model aims to identify and describe the face using more details with a minimum of five facial landmarks' positions.

2. After the first step of facial recognition is complete, we need to utilize a library known as seg-mask motifs, which is used for accurate eye segmentation. This library, as demonstrated in the Github repository described below, helps in performing semantic segmentation easily and internally uses a BiSeNet face segmentation for our task.

A. According to the GitHub description, the seg-mask motifs is “a package for easy generation of the binary semantic mask of different labels using multiple models easily. Moreover, it supports operations on the mask created for image editing.”
B. Bilateral Segmentation Network, also known as BiSeNet, works by designing a Spatial Path with a small stride to generate high-resolution features. Meanwhile, a Context Path with a fast-down sampling strategy is employed to obtain a sufficient receptive field. On top of both of those paths, a new Feature Fusion Module is introduced to combine features efficiently.
C.

Preliminary and Final Design

The following diagrams outline the preliminary designs and the final, along with the rationale.

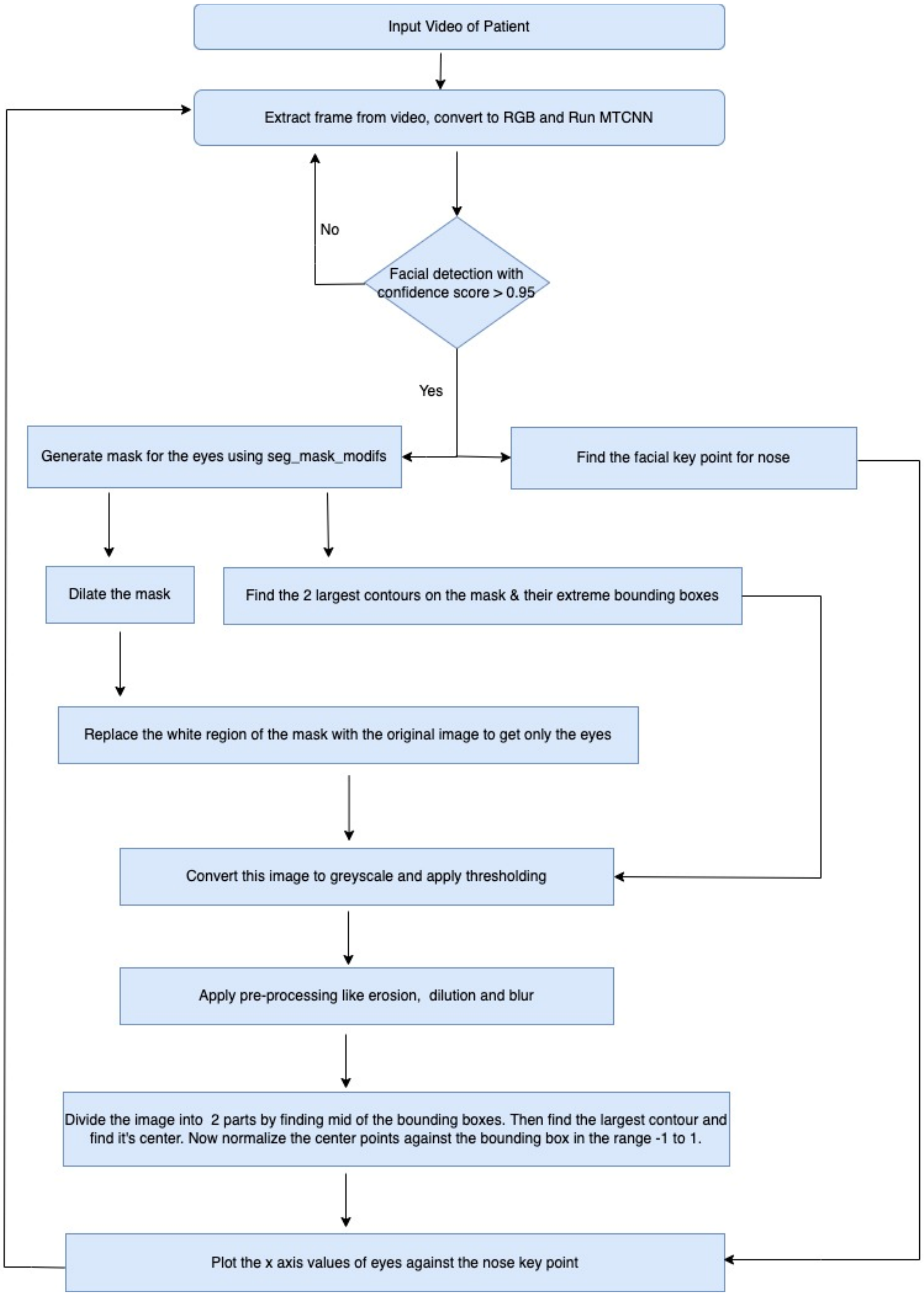

Input Video of Patient
Extract frame from video, convert to RGB and Run MTCNN
No
Facial detection with confidence score > 0.95
Yes
Generate mask for the eyes using seg_mask_modifs
Find the facial key point for nose
Dilate the mask
Find the 2 largest contours on the mask & their extreme bounding boxes
Replace the white region of the mask with the original image to get only the eyes
Convert this image to greyscale and apply thresholding
Apply pre-processing like erosion, dilution and blur
Divide the image into 2 parts by finding mid of the bounding boxes. Then find the largest contour and find it's center. Now normalize the center points against the bounding box in the range -1 to 1.
Plot the x axis values of eyes against the nose key point

Figure 2. This flowchart shows the procedure of accurately detecting saccades in a patient. Our procedure heavily depends on the MTCNN pre-trained model for facial detection as it is very accurate and robust despite different sizes, lighting, and strong rotations. On top of the MTCNN model, we added an algorithm to record the direction of the eye gaze in relation to the nose position and generate a graph.

The flowchart above helps us to understand the necessary steps we need to take to ensure the accuracy of this project. The crucial steps of that flowchart are listed below:

1. Extract the patient's face from the original video: This step is arguably the most important one because we need to identify the entire face region in order to locate the eye contours.

2. Find the 2 largest contours on the mask & their extreme bounding box points: As mentioned before, dlib's 68 facial landmarks are rendered useless when the patient's head is tilted at an angle and/or the ophthalmologist puts their hands on the patient.

   A. Create a mask of the eyes in the color white using seg_mask_modifs
   B. Divide the face and identify the two largest contours on both sides.
   C. Map the actual face on top of the white region with the eye regions still being detected

3. Apply pre-processing techniques such as erosion, blur, and dilation: These techniques are needed because we want to eliminate all the facial "holes" present on the face's mask.
4. Plot values of the x-axis of both eyes compared to the facial key point of the nose.

The steps above are critical to this success. The following figure will use pictures to illustrate why these steps are so crucial.

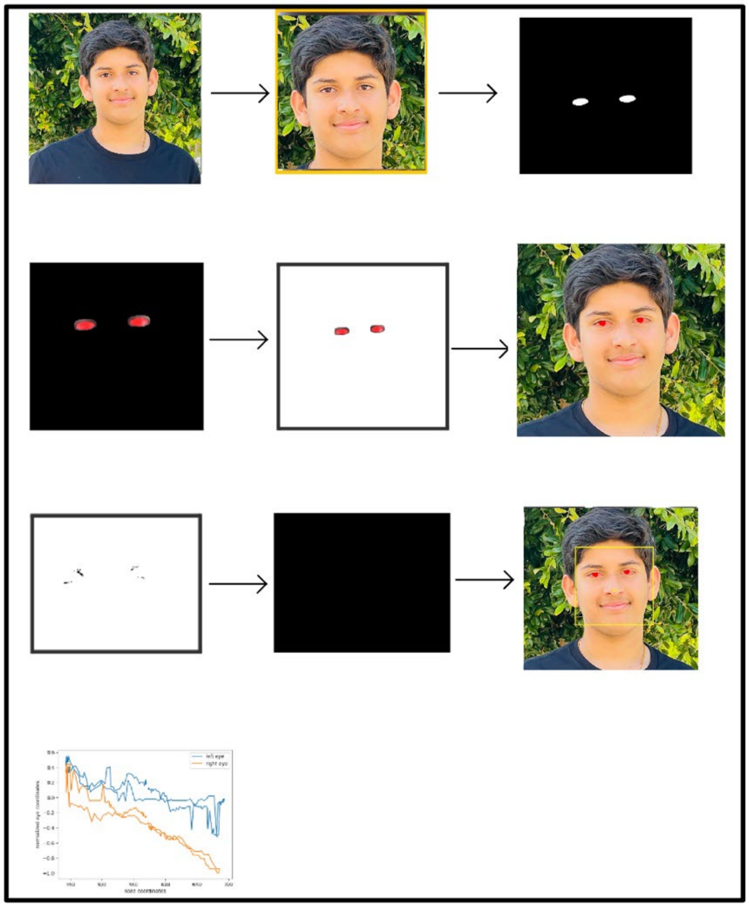


Figure 3. Algorithm Illustration

Prototype

Materials:

- Python 3.0
- Opencv
- MTCNN model
- TensorFlow
- Matplotlib
- NumPy
- Seg-mask-modifs including Pytorch, Torchvision, Pillow
- *Data recorded from online YouTube videos

The prototype was built using the designs shown in Figure 4, which requires inputs of recorded videos, with testing variables for optimization being facial detection using the MTCNN algorithm, segmentation using the BiSeNet face model, and facial landmarks as shown in Figure 1. Section 4, which includes Design Criteria, Testing, and Evaluation, highlights these independent variables in more detail.

The data needed for training and testing has to follow certain criteria:

- Patient’s entire face needs to be on the recorded video in at least 10 frames.
- Patient must perform the Head Impulse Test(HIT) in the recorded video of at least 1 turn.
- The Head Impulse Test is a test in which the patient’s face will be moved, while the patient tries to fixate his/her gaze on a particular point.
- Brightness must be at least 5 lux in order for the MTCNN model to detect the face properly.
- Video ideally must be between 30 seconds and 3 minutes.

Initial Testing, And Evaluation

The original algorithm to detect abnormal saccades heavily relied on the accuracy of the sixty-eight facial landmarks, as seen in Figure 1 above. The sixty-eight facial landmarks allow machine learning scientists to efficiently locate the facial landmarks, including the eyes. What makes these facial landmarks particularly more beneficial is the fact that the dlib library (the library that allows us to get the facial landmarks) is both reliable and fast, which is why it is the most popular facial landmark detector in computer vision libraries according to PyImageSearch University.

Despite the many advantages of dlib’s facial landmarks, when the face is occluded or at an angle to the camera, most facial landmarks appear to be disoriented. This problem further increases when a doctor performing the HIT test on a patient records the video with her hands on the patient’s face. This is the case in the two figures below.

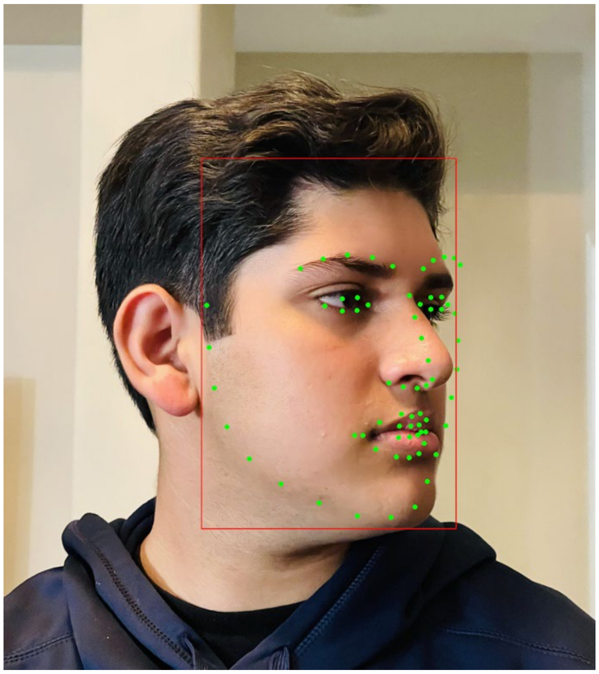


Figure 4 (a) Facial Landmark detection when face turned left

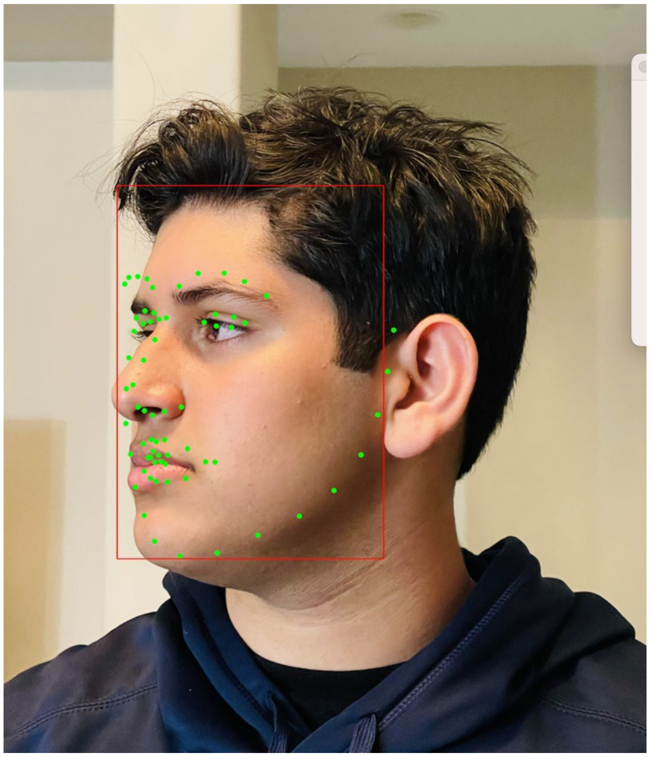


Figure 4 (b) Facial Landmark detection when face turned right

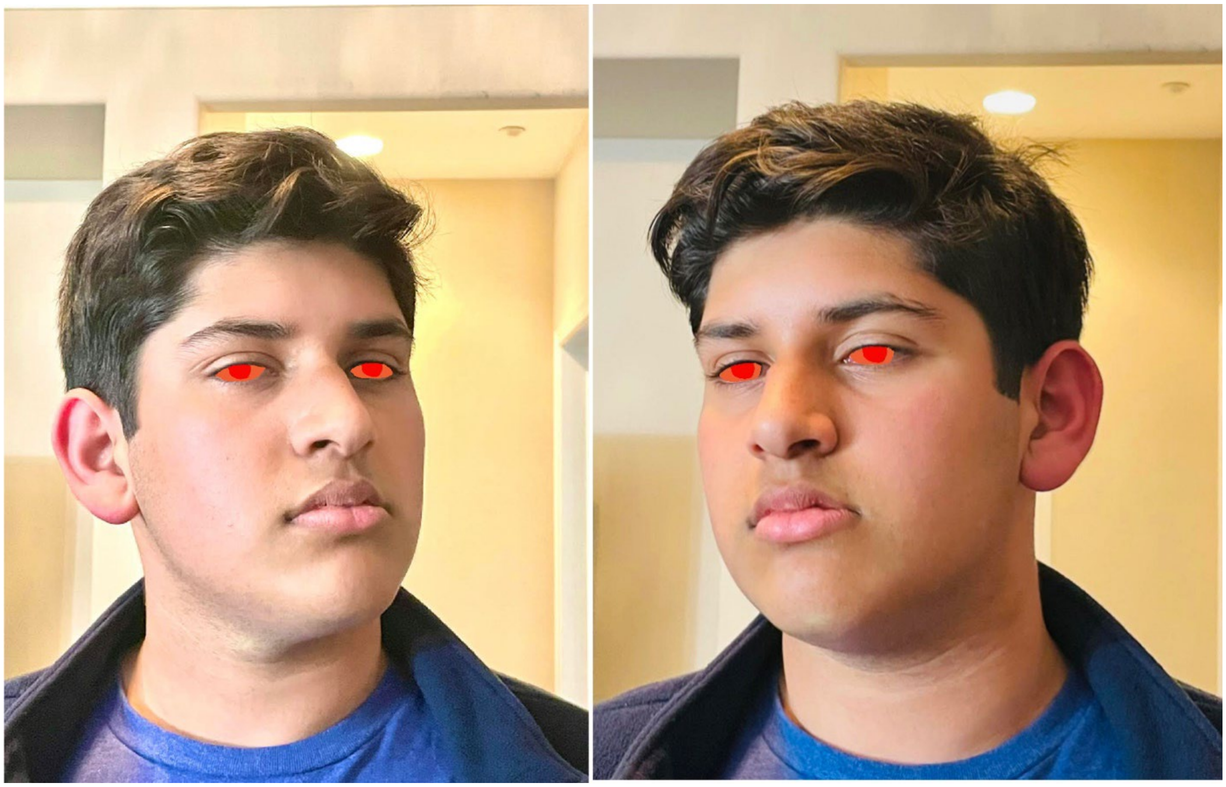
Redesign, Testing, and Evaluation

To resolve this issue, I decided to use the MTCNN algorithm as it is very accurate and robust. It properly detects faces even with different sizes, lighting, and strong rotations. MTCNN is as lightweight a solution as possible can be. We will construct an MTCNN detector first and feed a NumPy array as input to the detect faces function under its interface. I load the input image with OpenCV in the following code block. The detect faces function returns an array of objects for detected faces. The returned object stores the coordinates of detected faces in the box key. The main purpose of the MTCNN algorithm in this project is the following:

1. Detect the face from the frame robustly.

2. Identify the nose key point to make the x-axis the nose coordinates. Using these nose coordinates, the graph will allow us to simultaneously compare the x-coordinates of both eyes.

3. After implementing the MTCNN detector, the BiSeNet model has a function called seg_mask_modifs, which is used to create the facial mask. As outlined in the preliminary and final design, we identify the two largest contours on both sides of the face in order to get the accurate contour region(in this case, the two eyes). By extracting the eye regions, we are able to get one step closer to solving the issue of accurately detecting abnormal saccades. After extracting the largest contours, we apply erosion, dilation, and blur features to remove the “broken parts” of the eyes or the holes. By using the outlined procedure, I generated an image below which precisely detects the contours of both eyes, shown in red.



Figures 5. As you can see through these two images, even though the patient's head is tilted at a certain angle, the model, unlike the 68 facial landmarks, can still detect the eye regions effectively (barring some disadvantages).

Results

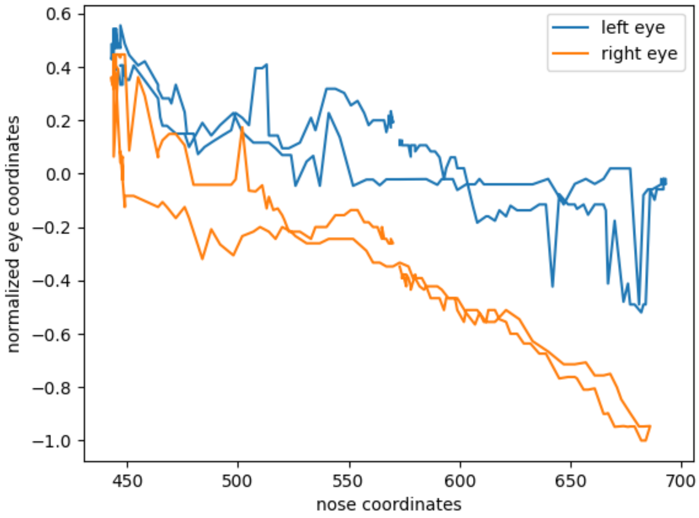


Figure 6 (a) Graph representing regular saccades

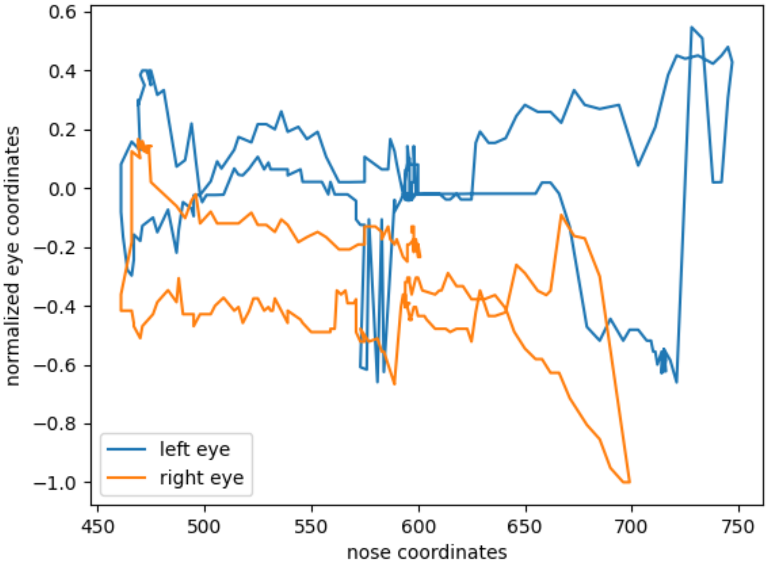


Figure 6 (b) Graph representing abnormal saccades

Figure 6(a) and 6 (b). These figures show both regular saccades and abnormal saccades. Since the graph of the first image resembles that of a y = -x linear graph and both eye regions have little to no fluctuations, we can conclude that a patient does not have saccades. However, in the second figure, there are more drastic fluctuations, which is why one can conclude that the subject has abnormal saccades.

Conclusion

It is hard to diagnose if someone has abnormal saccades. Despite advances in the field of detecting saccades such as vHIT (Video Head Impulse Test) and the newly-implemented cHIT (Clinical Head Impulse Test), these tests may be inaccurate as well as time-consuming for ophthalmologists. To solve this problem, machine learning scientists have created eye gaze estimation algorithms such as RT-GENE and iTracker, but they are not precise and are expensive, which is why the problem of detecting abnormal saccades is so intriguing.

To fix this, the experimenter attempts to develop an algorithm that can accurately detect both pupils and create a graph showing the movements of both eyes' gazes. The hypothesis was that the new algorithm as outlined will outperform in its ability to accurately diagnose abnormal saccades and its speed compared to ophthalmologists performing the HIT test on a patient. The project demonstrated this as our model generated a precise graph of the eye movements, allowing one to easily detect saccades rather than wasting expensive resources. What makes this model even more exciting is that the experimenter just needs to satisfy the basic requirements and does not need any expensive equipment. By fulfilling these requirements, an accurate judgment can be made from the graph.

The final design does have some limitations. For example, we assume that while the patient is performing the HIT test, they are not vertically moving their head. In other words, no y-axis movement is allowed. Another constraint of this model is that it does not output the binary value of abnormal saccades. Instead, one has to identify if a patient has abnormal saccades by analyzing the graph. Finally, few datasets are publicly available, making it harder to test how accurate this model is.

Acknowledgements

I would like to thanks my mentor, Isabel Hyo Jung Song, Professor, San Francisco State University.

References

Jay, Doctor. “Northoak Chiropractic.” Northoak Chiropractic, Northoak Chiropractic, 5 May 2022, https://northoakchiro.com/blog/2022/03/14/saccades-brain-injuries/#:~:text=Saccades%20are%20eye%20movements%20that,of%20gaze%20fixation%20to%20another .

Danaadmin. “The Eyes Are Windows into the Brain.” Dana Foundation, Dana Foundation, 7 Aug. 2019, https://dana.org/article/the-eyes-are-windows-into-the-brain/.

Kaspersky. “What Is Facial Recognition – Definition and Explanation.” Www.kaspersky.com, 9 Feb. 2022, https://www.kaspersky.com/resource-center/definitions/what-is-facial-recognition .

Halmagyi, G. M., et al. “The Video Head Impulse Test.” Frontiers, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/fneur.2017.00258/full .

Kerber, Kevin A. “Vertigo and Dizziness in the Emergency Department.” Emergency Medicine Clinics of North America, U.S. National Library of Medicine, Feb. 2009, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2676794/ .

Anto, et al. “Facial Landmarks with Dlib, Opencv, and Python.” PyImageSearch, 3 July 2021, https://pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/ .

“Measuring Saccade Latency Using Smartphone Cameras.” IEEE Xplore, https://ieeexplore.ieee.org/document/8703178.

JD;, Wong EC;Pasquesi L;Steenerson KK;Sharon. “A Broader View of Video Head Impulse Tests-Reframing Windows.” JAMA Otolaryngology-- Head & Neck Surgery, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/33270083/.

IS;, Halmagyi GM;Curthoys. “A Clinical Sign of Canal Paresis.” Archives of Neurology, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/3390028/ .

Mantokoudis G;Saber Tehrani AS;Kattah JC;Eibenberger K;Guede CI;Zee DS;Newman-Toker DE; “Quantifying the Vestibulo-Ocular Reflex with Video-Oculography: Nature and Frequency of Artifacts.” Audiology & Neuro-Otology, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/25501133/ .

Contributor, TechTarget. “What Is Polynomial Interpolation? - Definition from Whatis.com.” WhatIs.com, TechTarget, 26 Apr. 2013, https://www.techtarget.com/whatis/definition/polynomial-interpolation#:~:text=Polynomial%20interpolation%20is%20a%20method,can%20be%20made%20by%20interpolation .

Janky KL;Patterson JN;Shepard NT;Thomas MLA;Honaker JA; “Effects of Device on Video Head Impulse Test (Vhit) Gain.” Journal of the American Academy of Audiology, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/28972467/#:~:text=There%20was%20not%20an%20effect,gain%20values%20between%20devices%2Falgorithms .

“Dizziness.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 15 Oct. 2020, https://www.mayoclinic.org/diseases-conditions/dizziness/symptoms-causes/syc-20371787#:~:text=Dizziness%20is%20a%20term%20used,reasons%20adults%20visit%20their%20doctors .

Suh MW;Park JH;Kang SI;Lim JH;Park MK;Kwon SK; “Effect of Goggle Slippage on the Video Head Impulse Test Outcome and Its Mechanisms.” Otology & Neurotology : Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/27956722/#:~:text=Results%3A%20The%20most%20common%20slippage,%2C%20and%204)%20deceleration%20bumps .

Agarwal, Vardan. “Real-Time Head Pose Estimation in Python.” Medium, Towards Data Science, 30 July 2020, https://towardsdatascience.com/real-time-head-pose-estimation-in-python-e52db1bc606a .

Palmero, Cristina, et al. “Recurrent CNN for 3D Gaze Estimation Using Appearance and Shape Cues.” ArXiv.org, 17 Sept. 2018, https://arxiv.org/abs/1805.03064.