Media Engineering and Technology Faculty
German University in Cairo

Indoor Scenes Understanding for Visual Prostheses

Bachelor Thesis

Author: Rawan Fouda
Supervisors: Dr. Seif Eldawlatly
Msc. Eng. Reham Elnabawy
Submission Date: 12 June, 2022

Media Engineering and Technology Faculty
German University in Cairo

Indoor Scenes Understanding for Visual Prostheses

Bachelor Thesis

Author: Rawan Fouda
Supervisors: Dr. Seif Eldawlatly
Msc. Eng. Reham Elnabawy
Submission Date: 12 June, 2022

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowledgement has been made in the text to all other material used

---

Author
12 June, 2022

Acknowledgments

I would like to thank several people for their help and support during the production of this thesis. I am extremely thankful to my supervisors, Dr. Seif Eldawlatly and MSc. Eng. Reham Elnabawy for their guidance and support and for providing me with such an opportunity to work on something as astounding as Visual Prostheses. I would also like to thank all of my friends and family members for encouraging and supporting me whenever I needed them. This work could not have been done without you.

Abstract

A visual prosthesis is a visual device that restores vision for patients with partial or total blindness. The aim of this project is to form a realistic simulated prosthetic vision and enhance prosthetic vision to allow patients to comprehend indoor scenes more easily and confidently. For this to happen, machine learning models and image processing techniques are applied to form a clearer image of the scene captured by the camera. In this project, different image processing techniques are applied and tested through experiments conducted on the screen and in virtual reality. The results show that the enhancement techniques do have an effect on the understanding of indoor scenes, with a greater percentage of room and object recognition in the Enhancement group. The p-value obtained using the t-test and ANOVA statistical tests was less than 0.05, indicating that the difference is significant. In the future, applying these enhancement techniques as an integral part of the visual prosthesis devices will give patients more ease in determining the room they are currently in and increase their awareness of their surroundings.

Contents

Chapter 1

Introduction

1.1 Motivation

Visual prostheses work by using images captured by a camera and then converting these images to electrical impulses to stimulate the remaining functional neurons in the visual pathway. Visual prostheses do not fully restore vision, and there remains a large difference between normal and prosthetic vision. Patients who use visual prostheses still face difficulty recognizing their surroundings, indicating that the output of visual prostheses remains a problem to be tackled. The goal of conducting research on visual prostheses is to restore vision to a greater degree in patients who have lost vision due to eye diseases or unfavourable circumstances. This would help them increase their awareness of their surroundings and make them more independent in their lives. Visual prosthesis continues to be a great area of research due to its tremendous benefits for blind people, and there have been several implementations of it in real life, showing some promising results.

1.2 Thesis Objective

The aim of this thesis is to develop a better phosphene representation of indoor scenes for people with some sort of blindness to help them comprehend indoor scene types more easily. Machine learning and image processing techniques are used as a part of image enhancement and a phosphene simulation library is used to predict the actual outcome given the input image.

1.3 Thesis Contributions

Despite Mask-RCNN being used in previous papers, this project introduces further enhancements to the output of the object segmentation by using image processing techniques using OpenCV instead of directly using a binary mask representation as in previous research [1, 2]. Furthermore, more experiments have been conducted using different techniques such as virtual reality and real-time virtual reality, in addition to the normal screen experiments, to test the validity of the final output and identify all the possible struggles that the patient might encounter, and also to improve the validity of the measurement of how the enhancement technique has improved vision for patients. Final results show that the enhancement approach, where machine learning and image processing techniques were involved, did in fact improve the scene understanding, as more participants were able to identify the type of room and the objects in it correctly. The number of participants who were able to identify the room type as well as the objects correctly was also significantly greater compared to the direct group, with a p-value less than 0.05.

1.4 Thesis Outline

This thesis is divided into five main sections. In Chapter 1, we have the introduction, which explains the general topic of the thesis and the aim of the thesis, in addition to the thesis contributions. In Chapter 2, there is the background, which provides information about the main topics covered in the thesis, including blindness, visual prostheses, machine learning, image processing techniques, and literature review. Chapter 3, which is the Methodology chapter, explains the implementation of the thesis, the phosphene simulation, the methods, and the experiment setup. Then we have Chapter 4, the results section, which analyses the results of the experiments and investigates the thesis's outcome. Finally, in Chapter 4, there is the conclusion and future work section, which includes the thesis's summary, discussion, and future work.

Chapter 2

Background

2.1 Blindness

According to the World Health Organization (WHO), an estimated 2.2 billion people suffer from some form of vision impairment [3] . Some of the leading causes include age-related macular degeneration, retinitis pigmentosa, glaucoma, and diabetic retinopathy.

2.1.1 Age-related macular degeneration

A group of conditions that cause damage to the centralized vision (macular) leads to age-related macular degeneration (AMD). There are two types of macular degeneration: dry AMD and wet AMD, as shown in Figure 2.1. More than a third of the people who have AMD have the dry form where parts of the macular get thinner and drusen (tiny protein clumps) grow [4]. The other form, wet AMD, is less common but much more severe. This happens when new, abnormal blood vessels grow under the retina. These vessels may leak blood or other fluids, causing scarring of the macula.

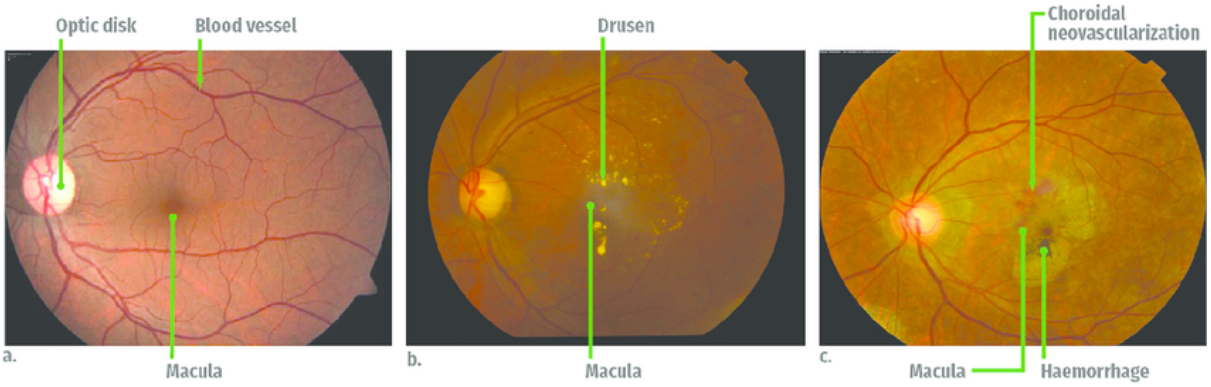


Figure 2.1: First subfigure is the normal form. The second figure shows dry AMD and the last one shows wet AMD [5]

Symptoms of macular degeneration include visual distortions where straight lines appear wavy, blurriness of vision, difficulty reading printed words and recognizing faces, and the need for brighter light due to difficulty adapting to low light levels.

There are several risk factors for AMD, such as age (people aged 50 and above are more susceptible to AMD) [6], smoking, high blood pressure, and obesity.

There are several tests that an eye doctor performs to diagnose patients with AMD [7]. The doctor can measure vision ability at various distances using a visual acuity test or perform a close-up examination of the retina by pupil dilation using eye drops. For wet AMD, diagnosis is conducted through Fluorescein angiography or Amsler grid. Fluorescein angiography involves a special dye being injected into a vein in the arm. Pictures are then taken as the dye passes through the blood vessels in the retina, helping the doctor to determine if the blood vessels are leaking. The Amsler Grid, as shown in Figure 2.2 uses a checkerboard-like grid to determine if the straight lines in the pattern appear bent or missing to the patient [8].

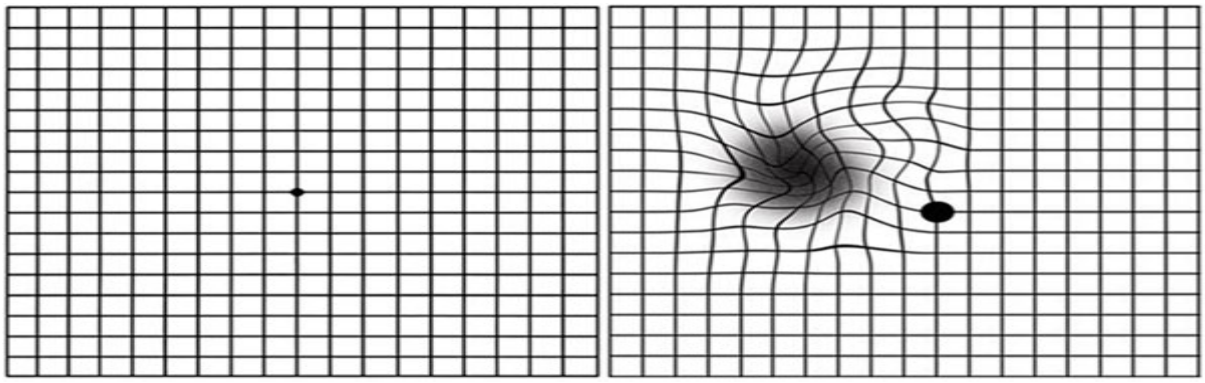


Figure 2.2: Second picture here shows how patients with AMD view the first picture [9]

There is no treatment for dry AMD, but visual aids can be used. Wet AMD may need regular eye injections and, sometimes, a light treatment, called photodynamic therapy, to preserve the quality of vision. AMD is often linked to an unhealthy lifestyle. Thus, it is advised to lose weight if you are obese, exercise, eat a balanced diet, and quit smoking.

2.1.2 Retinitis Pigmentosa

Retinitis Pigmentosa (RP) is a genetic eye disease that affects the retina. It is one of the leading causes of blindness [10]. As shown in Figure 2.3, RP makes cells in the retina degenerate over time, causing loss in vision.

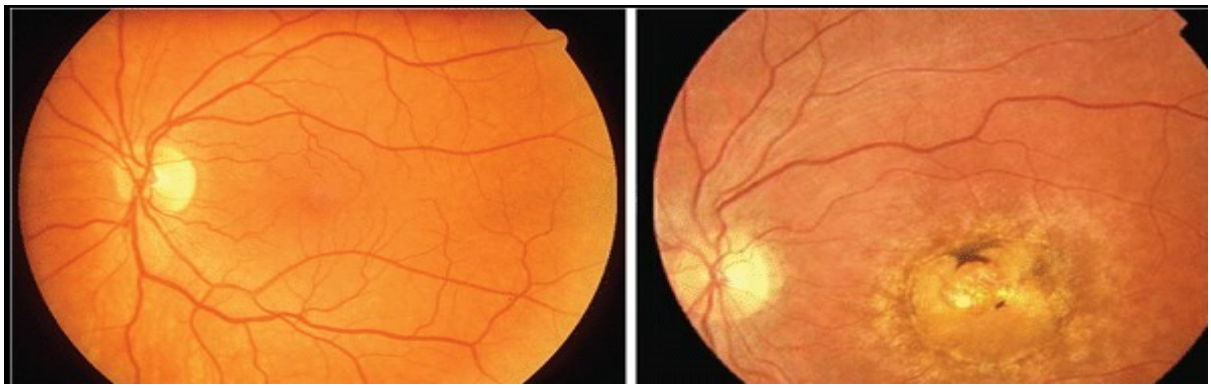


Figure 2.3: At the left, the normal retina; right, Retinitis Pigmentosa with retinal hyperpigmentation in a characteristic pattern [11]

Symptoms of RP include blindness to colour and sensitivity to bright light, difficulty reading print and figuring out detailed images, and a hard time stumbling over objects that are not seen. RP patients also tend to glare a lot. The main risk factor is a family history of Retinitis Pigmentosa.

Retinitis Pigmentosa can be diagnosed through pupil dilation using eye droplets. Other tests include electroretinography (ERG), where the doctor checks how well your retina responds to light; Optical Coherence Tomography (OCT) test [12] which uses light waves to take a detailed picture of your retina; fundus autofluorescence (FAF) imaging [13], which uses blue light to take a picture of the retina; and genetic testing to learn about the type of RP they have.

There is no treatment for RP, but low vision aids and rehabilitation (training) programs can help people with RP to restore some of their vision and make the most of it.

2.1.3 Glaucoma

Glaucoma is an eye disease where the optic nerve gets damaged. This is caused when fluid builds up and increases pressure inside the eye, as shown in Figure 2.4. Glaucoma is the second leading cause of blindness in the world [14]. If it's not diagnosed or treated early, it may lead to loss of vision.

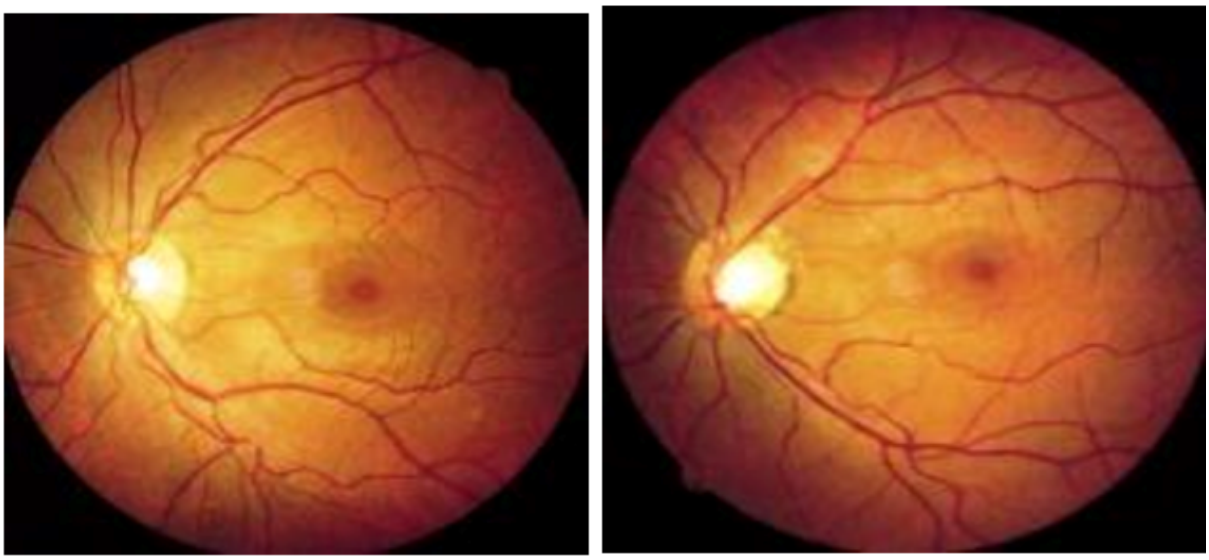


Figure 2.4: An example of visually presented glaucoma: (a) normal retinal image, (b) glaucoma [15]

Risk factors for glaucoma include having high internal eye pressure, diabetes, heart disease, high blood pressure, having corneas that are thinner in the centre and being over 60.

There are several types of glaucoma: open-angle, closed-angle, congenital, and normal-tension [16]. Open-angle is the most common type that happens when tiny deposits build up in the eye-drainage canals, causing fluid build-up and pressure on the optic nerves. Closed-angle, which is a rare type, happens when the angle between the iris and the cornea is narrow, causing the drainage canals in the eye to become blocked and higher fluid pressure. Some newborns have congenital glaucoma, so it could be genetic. There is a fourth type, normal tension, which is commonly found in Asians and Asian Americans and has no scientific reason. Generally, glaucoma affects both eyes. There are several tests conducted by an eye doctor to diagnose glaucoma. Usually, the eye doctor may do several of these tests. Pupil dilations are used to enlarge the pupils and view the optic nerve at the back of the eyes. There is also the slit-lamp exam to examine the inside of the eye. Gonioscopy is used to examine the angle where the iris and cornea meet [17]. Pachymetry is used to measure the thickness of the cornea [18]. Optical coherence tomography is used to check for changes in the optic nerve [19]. The ocular pressure test is used to measure eye pressure [20]. Visual acuity test (eye charts) to check for vision loss and a visual field test (perimeter) to check for changes in peripheral vision [21].

Damage caused by glaucoma cannot be reversed, but it can be treated by lowering the eye pressure. Depending on the severity of the condition, glaucoma can be treated via eye drops, oral medication, or surgery. There are several eye drop medications that can be used, which include Prostaglandins, Beta blockers, Alpha-adrenergic agonists,

carbonic anhydrase inhibitors, Rho-kinase inhibitors, and Miotic or cholinergic agents. Prostaglandins increase and Miotic or cholinergic agents increase the outflow of the fluid in your eye. Beta-blockers and carbonic anhydrase inhibitors reduce the production of fluid in your eye [22]. Alpha-adrenergic agonists such as apraclonidine (Iopidine) and brimonidine (Alphagan P, Qoliana) reduce the production of aqueous humor and increase outflow of the fluid in your eye [23]. Rho-kinase inhibitors work by suppressing the rho kinase enzymes responsible for fluid increase [24]. All of these eye drop medications may have some side effects unrelated to the eyes. In addition to eye-drop medications, oral medications may be used. Another invasive approach is surgery. If the patient has open-angle glaucoma, the eye doctor may resort to laser therapy, which uses a small laser beam to open clogged channels in the trabecular meshwork [25]. Filtering surgery, where the surgeon creates an opening in the sclera and removes part of the trabecular meshwork, [26]. Drainage tubes involve inserting a small tube shunt in your eye to drain away excess fluid to lower your eye pressure [27]. Alternative medicine includes herbal remedies, relaxation techniques, and Marijuana [28]. People with glaucoma need to be consistent with their medications, eat a healthy diet, exercise safely, and limit their caffeine intake to help limit the severity of glaucoma.

2.1.4 Diabetic retinopathy

Diabetic retinopathy is an eye disease caused by diabetes due to high blood pressure that damages the retina [29].

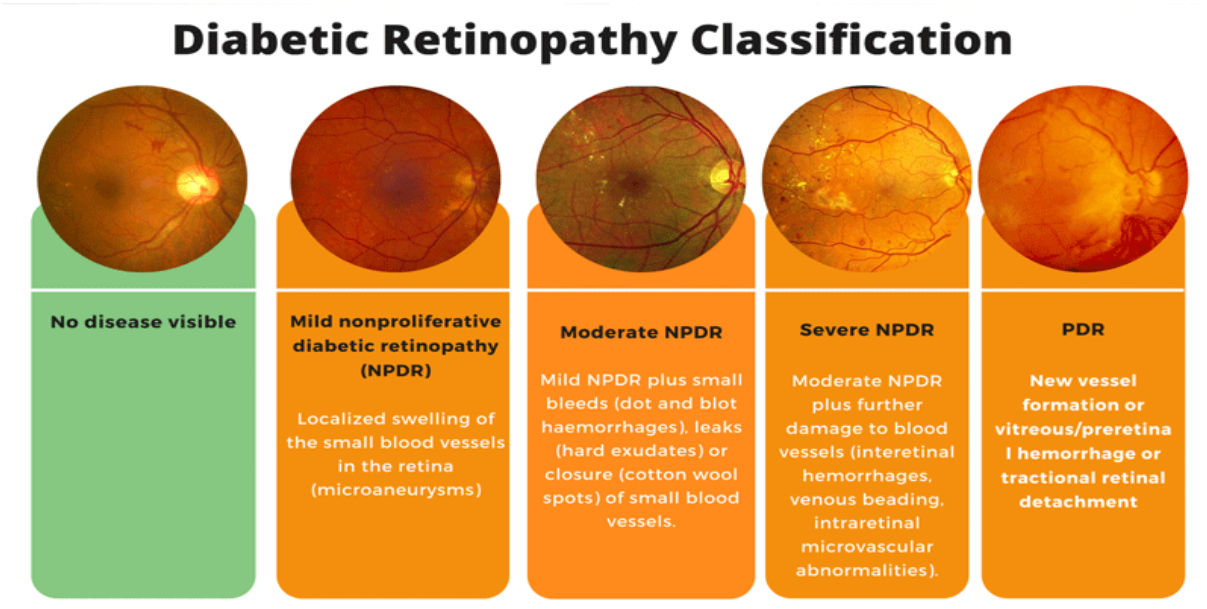


Figure 2.5: Stages of diabetic retinopathy [30]

Risk factors for developing diabetic retinopathy include having diabetes for a long time, poor control of blood sugar level, high blood pressure, high cholesterol, pregnancy, and being black/Hispanic/Native American [31].

Diabetic retinopathy has 4 stages, as shown in Figure 2.5. The first stage is mild nonproliferative diabetic retinopathy, where tiny areas of swelling appear in the retinal blood vessels. The second stage is moderate nonproliferative diabetic retinopathy, characterised by the increased swelling interfering with blood flow to the retina. The third stage is severe nonproliferative diabetic retinopathy, where the body receives signals to start growing new blood vessels in the retina, and the fourth stage is proliferative diabetic retinopathy, in which new blood vessels form in the retina.

Symptoms of diabetic retinopathy include an increased number of eye floaters, blurry vision, eye pain or redness, distorted vision, and poor night vision [32]. Diabetic retinopathy requires a comprehensive eye examination that involves measuring visual acuity, eye muscle movement, peripheral vision, depth perception and curvature of the cornea. Several treatments are available for diabetic retinopathy. Laser treatment, also known as photocoagulation [33], reduces the drive for abnormal blood vessels and swelling in the retina. Eye medications such as anti-VEGF medication, which is a steroid injection, can reduce swelling in the macula and improve vision [34]. If patients have diabetic retinopathy, they might need an eye surgery called vitrectomy [35], which is an operation to remove blood or scar tissue from the eyes.

A person with diabetic retinopathy needs to maintain their blood sugar, cholesterol, and blood pressure at target levels, get their eyes screened, and take diabetic medication. Furthermore, it is essential to maintain a healthy weight, eat a balanced diet, and stop smoking.

2.2 Visual Prosthesis

Visual prostheses serve as a lifeline to help people with blindness to restore their vision if no alternative medical treatment exists. This is especially important to allow patients to regain their independence and mobility.

2.2.1 Visual Prosthesis types

To understand how visual prostheses work, we need to understand the visual pathway, as shown in Figure 2.6.

First, light enters through the cornea. From the cornea, light passes through the pupil. The iris controls the amount of light that enters the pupil. Then it hits the lens, which focuses the light onto the retina. The light then passes through the vitreous humor and reaches the retina. The optic nerve is then responsible for transmitting the signals to the visual cortex. First, light enters through the cornea. From the cornea, light passes through the pupil. The iris controls the amount of light that enters the pupil. Then it hits the lens, which focuses the light onto the retina. The light then passes through the vitreous humor and reaches the retina. The optic nerve is then responsible for transmitting the signals to the visual cortex.

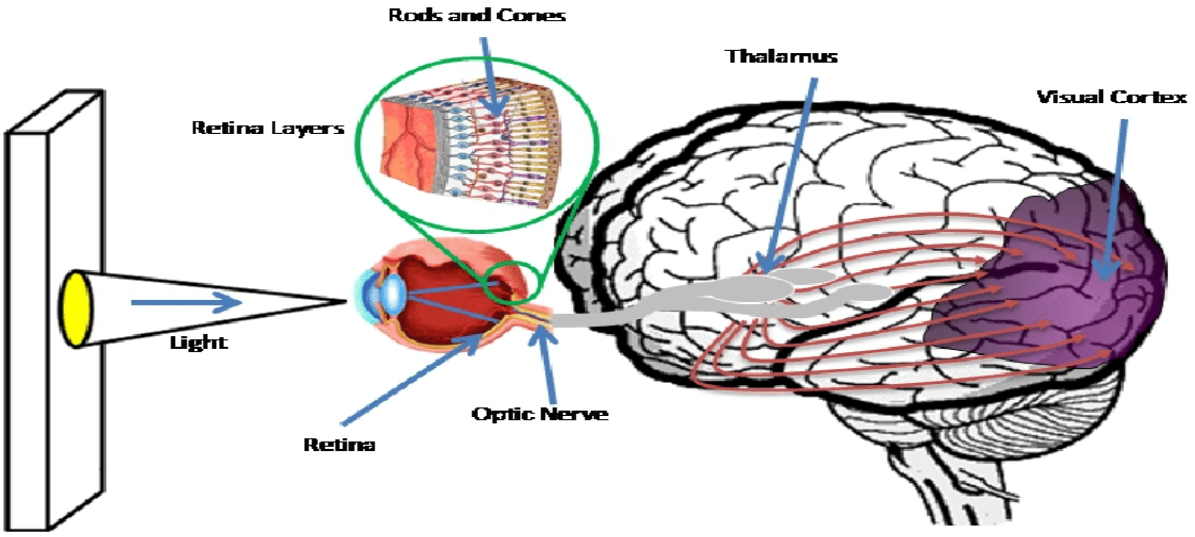


Figure 2.6: Human Visual Pathway [36]

There are three main types of visual prosthesis: retinal implant, optic nerve implant, cortical implant, and thalamic implant.

A retinal implant is a type of implant where electrodes stimulate the retinal cells based on the light patterns detected in the field of vision [37]. The resulting vision is a set of phosphenes displayed in a restricted portion of the visual field. The retinal stimulating type is the subject of most research because it is easier to access by surgery than any other target tissue for electrical stimulation.

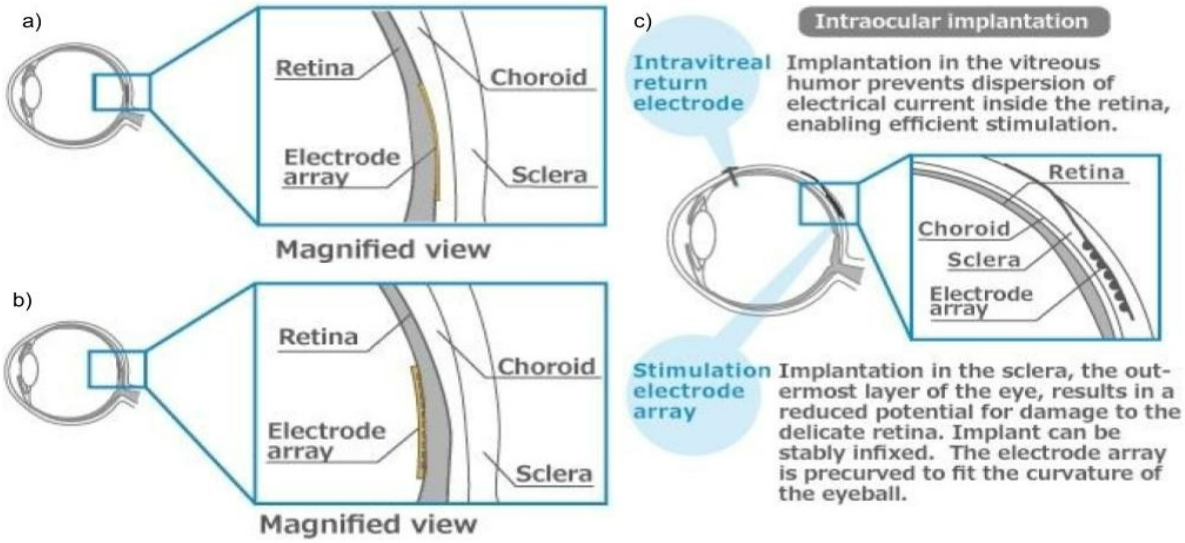


Figure 2.7: Set up of retinal implants: (a) epiretinal, (b) subretinal, (c) STS [38]

There are several types of retinal implants, as shown in Figure 2.7:

1. Epiretinal implant
   An electrode array is implanted on the retinal surface. The most common types are Argus I and Argus II, as shown in Figure 2.8. Argus II is the newer version, which has 60 electrodes arranged in a 6×10 format.

2. Subretinal Implant
   An electrode array is implanted under the retina. Alpha IMS is a subretinal implant with a high density of electrodes. However, it has a less robust packaging system. Alpha IMS electrodes are either 50×50 m or 100×100 m.

3. Suprachoroidal Transretinal Stimulation (STS)
   An electrode array is set up inside the sclera and a return electrode is set up inside the eye. This implant is considered to be safer and less invasive.

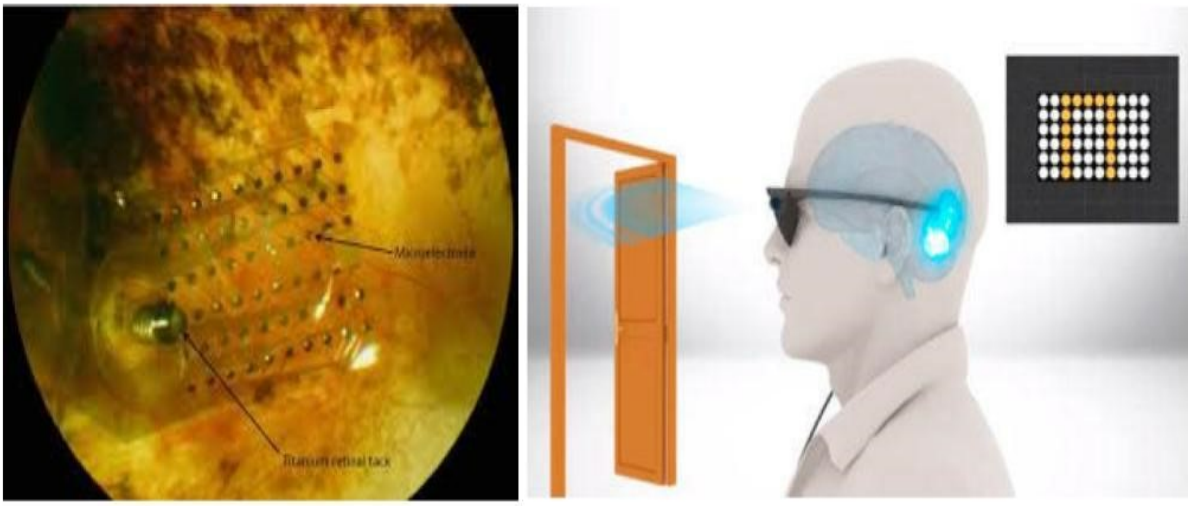


Figure 2.8: The Argus II retinal Prosthesis system [39]

In an optic nerve implant, a cuff electrode array, as shown in Figure 2.9, is wrapped around the optic nerves from the outside, and a type of stimulating optic nerve head with wire electrodes is being used [40].

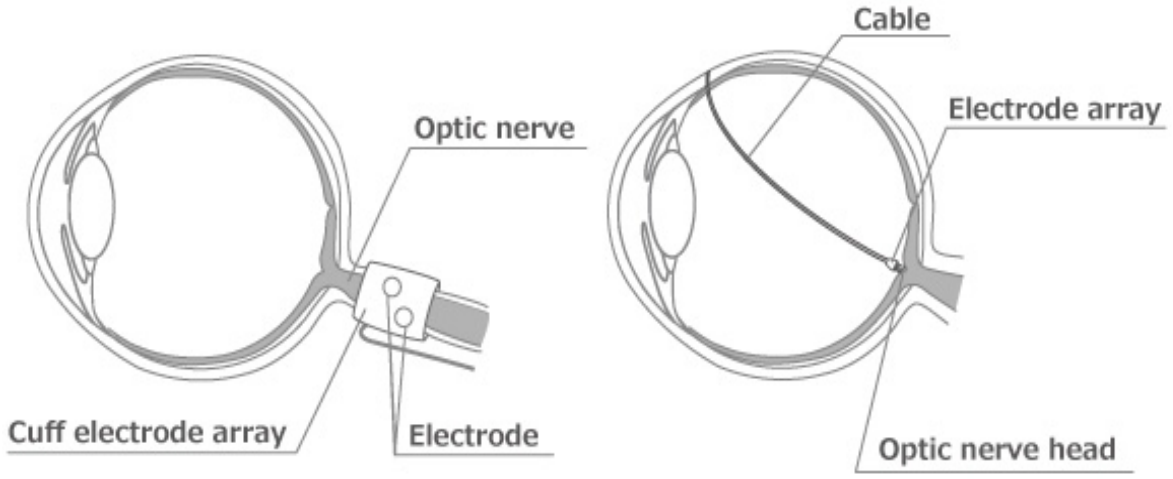


Figure 2.9: Optic Nerve Implant [38]

In a cortical implant, an electrode array is implanted in the brain, as shown in Figure 2.10, in the part of the visual cortex to produce vision [41]. However, this is a more difficult surgery and has lots of risks before and after the surgery.

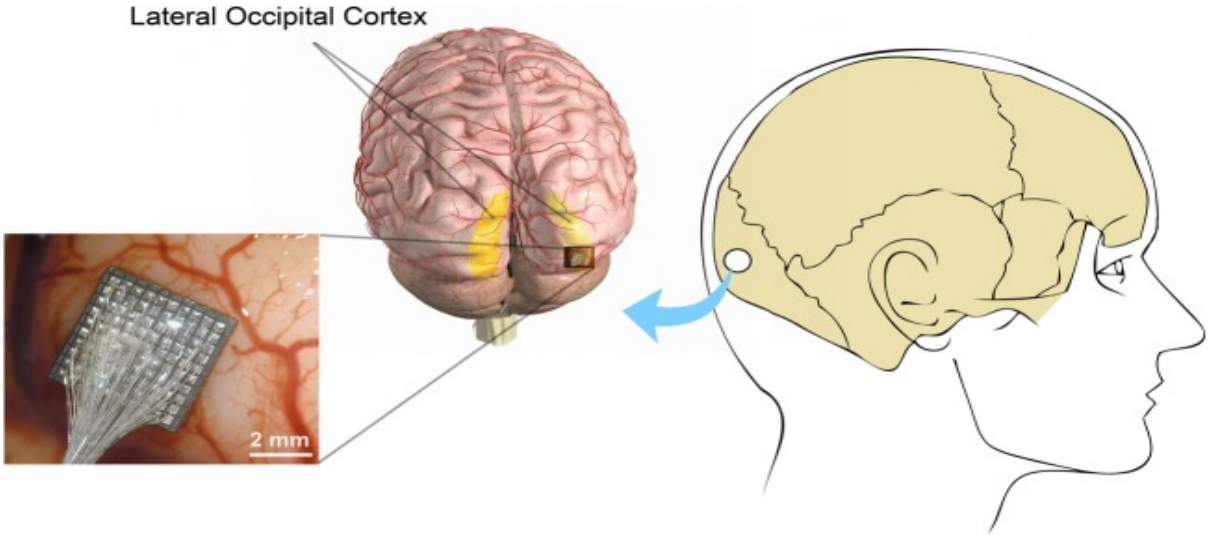


Figure 2.10: Cortical Implant for visual prosthesis [42]

In a thalamic implant, multiple microelectrodes are placed in the lateral geniculate nucleus of the thalamus, as shown in Figure 2.11, which is a part of the brain that relays signals from the eye to the visual system [43].

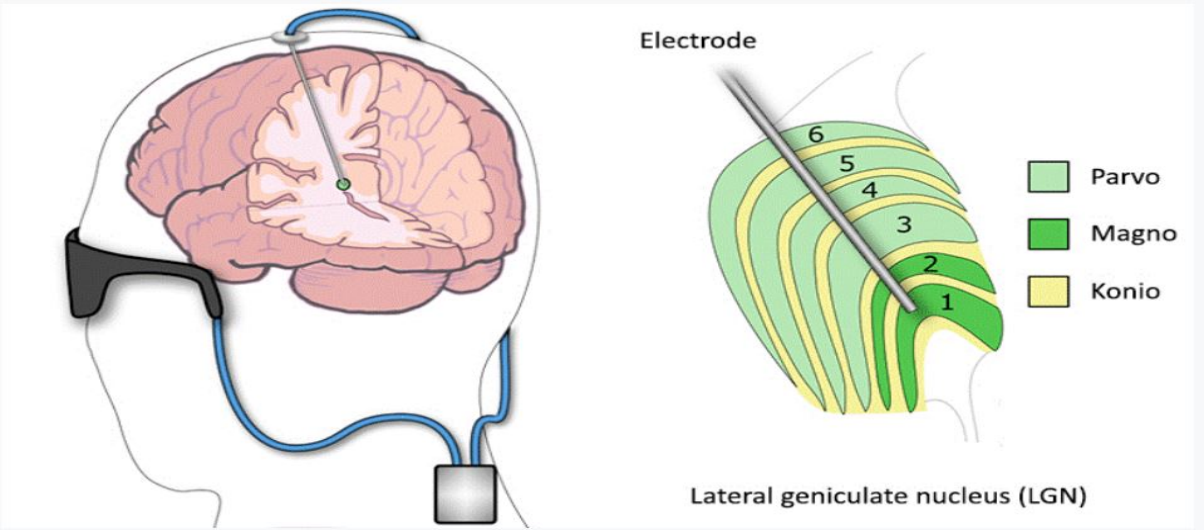


Figure 2.11: Thalamic Implant for visual prosthesis [44]

2.2.2 Simulated Prosthetic Vision

Prosthetic vision is created through the appearance of phosphenes. Phosphenes are produced from the inside of the eye more than from the external environment [45]. Seeing phosphenes occurs when a source (other than light) stimulates the retina or the optic nerve.

There are several representations of phosphenes in visual prosthesis.

Table 2.1: Summary of the appearances of phosphenes elicited via electrical stimulation at various sites in chronic human trials of vision prosthesis devices [46]

	Cortical	**Optic nerve**	**Epiretinal**	**Subretinal**
Shape	Round, diffused, matchstick, lines, square	Single patch, multiple dots, lines, triangles, colored background	Round, donut, line, cluster of dots	Round
Size	Punctuate (5 arc min), 1–2°, large coin (2.5°)	8–42 arc min, area of 1–50° squared	0.4–2°	5–30 arc min
Color	Colorless (white), yellow, gray, blue, red, brown, orange	White, yellow, blue, red	Yellow, white, green, blue, red–orange	Yellow, grayish
Visual location	Pseudo corticotopic	Depends on stimulation	Pseudo retinotopic	Retinotopic (perception of aligned phosphenes)
Flicker fusion	Variable, not always achieved, 20 Hz	8–10 Hz	40–50 Hz	Not reported
Multiplicity	Reflection about the horizontal and vertical meridians	Multiple dots, sometimes within colors background	Singular	Singular
Others	Depth perception	Moving phosphenes		
Modulation	Brightness, size	Brightness	Brightness, size	Brightness, size
References	[47, 48, 49, 50, 51]	[52, 53, 54]	[55, 56]	[57, 58]

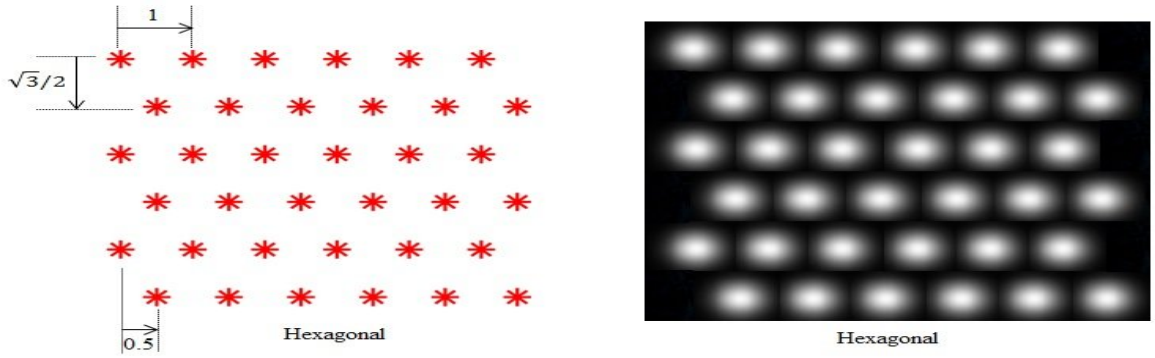


Figure 2.12: Phosphene Simulation [59]

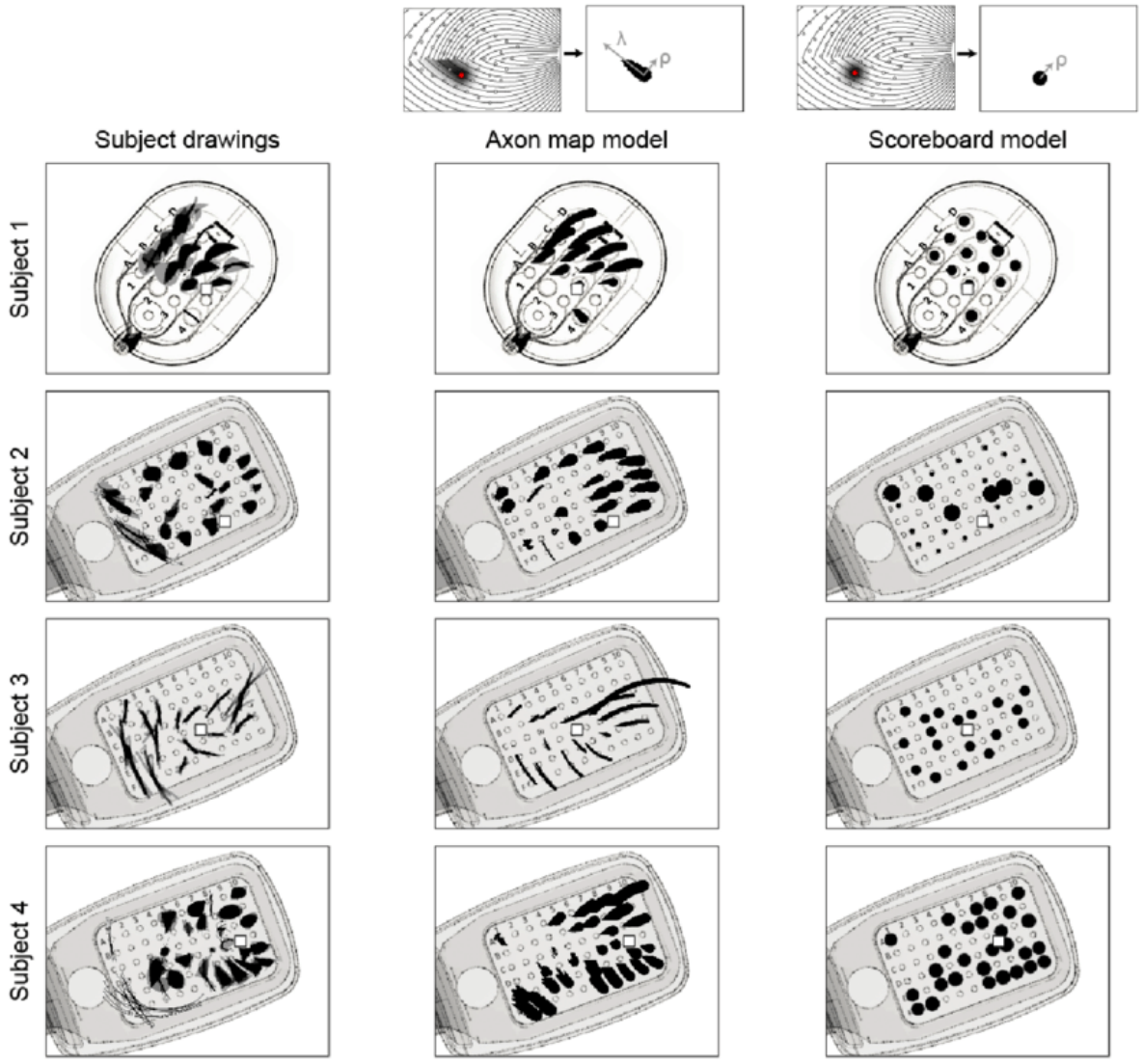


Figure 2.13: Phosphene drawings (left columns) contrasted against cross-validated phosphene predictions of the axon map model (center column) and the scoreboard model (right column) [60]

Figure 2.13 shows the different representations of phosphene simulations. The scoreboard model, as shown in Figure 2.12 assumes that electrical stimulation led to the percept of focal dots of light, centered over the visual field location associated with the stimulated retinal field location, whose spatial intensity profile decayed with a Gaussian profile.

Axonal stimulation, on the other hand, could lead to phosphenes that are elongated in shape but poorly localized.

2.3 Machine Learning and Image Processing techniques

Object segmentation is the process of extracting regions of interest from an image. There are several object detection algorithms, including YOLO, RCNN, and Mask R-CNN.

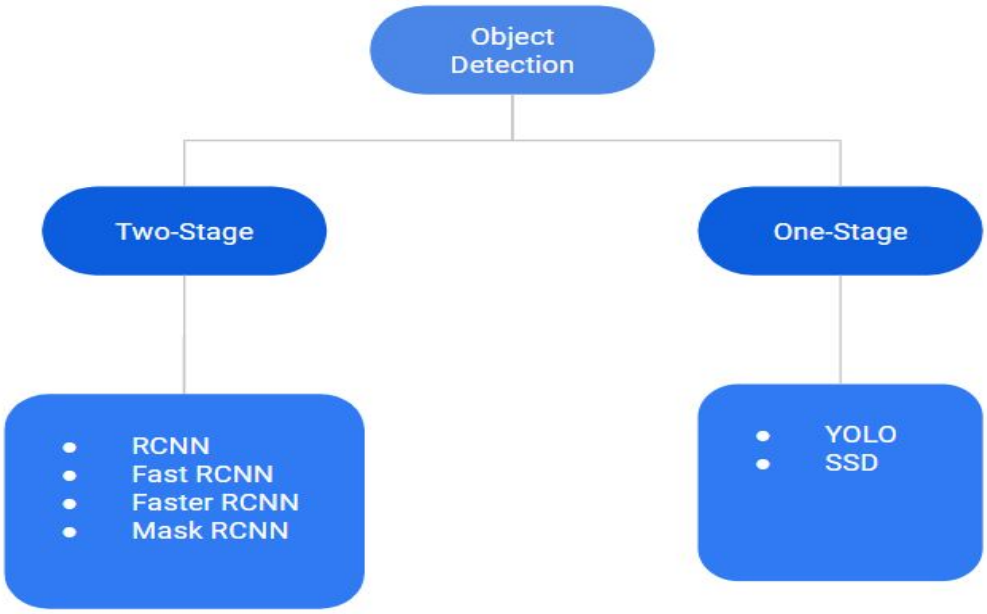


Figure 2.14: Examples of object detectors

There are two types of object detectors, as shown in Figure 2.14:

1. Two-Stage detectors
   In the first stage, region proposals are extracted using two networks: the backbone (for example, ResNet and VGG) and the region proposal network.

   In the second stage, objects are classified for each region's proposal. Examples include RCNN, Fast-RCNN, Faster-RCNN, and Mask-RCNN.

2. One-Stage detectors
   Object classification and bounding-box regression are done directly without using pre-generated region proposals (candidate object bounding-boxes). Examples include YOLO and SSD.

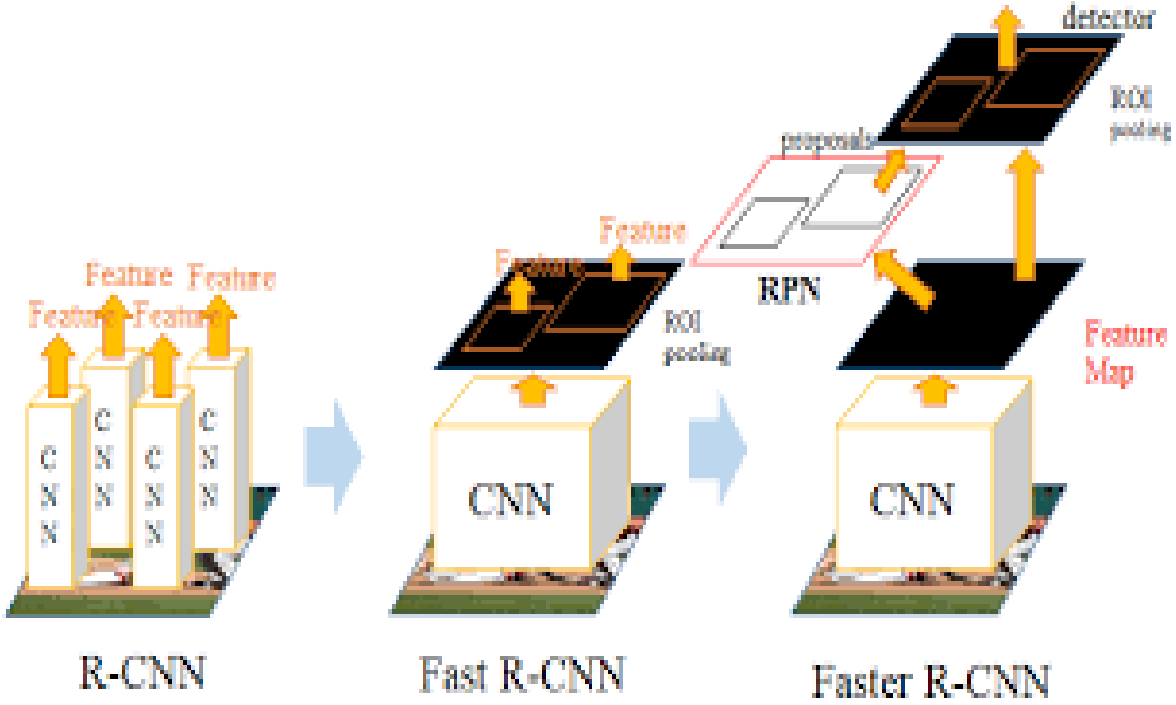


Figure 2.15: RCNN, Fast RCNN and Faster-RCNN [61]

For two-stage detectors, the newest version is Mask-RCNN, which has evolved from basic RCNNs, which is shown in Figure 2.15:

The R-CNN detector first generates region proposals using an algorithm like Edge Boxes. The proposal regions are extracted from the image and resized. Then, the CNN classifies the cropped and resized regions. Finally, the region proposal bounding boxes are refined by a support vector machine (SVM) [62].

Just like the R-CNN detector, the Fast R-CNN detector also uses an algorithm like Edge Boxes to generate region proposals. Unlike the R-CNN detector, which crops and resizes region proposals, the Fast R-CNN detector processes the entire image. Also, Fast R-CNN pools CNN features corresponding to each region's proposal instead of classifying each region [63]. Fast R-CNN is more efficient than R-CNN because the computations for overlapping regions are shared in Fast-RCNN.

The Faster R-CNN detector adds a region proposal network (RPN) to create region proposals directly in the network instead of using an external algorithm like Edge Boxes. Generating region proposals in the network is faster and better tuned to the input data [64].

Mask R-CNN is an extension of Faster R-CNN with an additional branch for predicting segmentation masks on each Region of Interest (RoI), as shown in Figure 2.16.

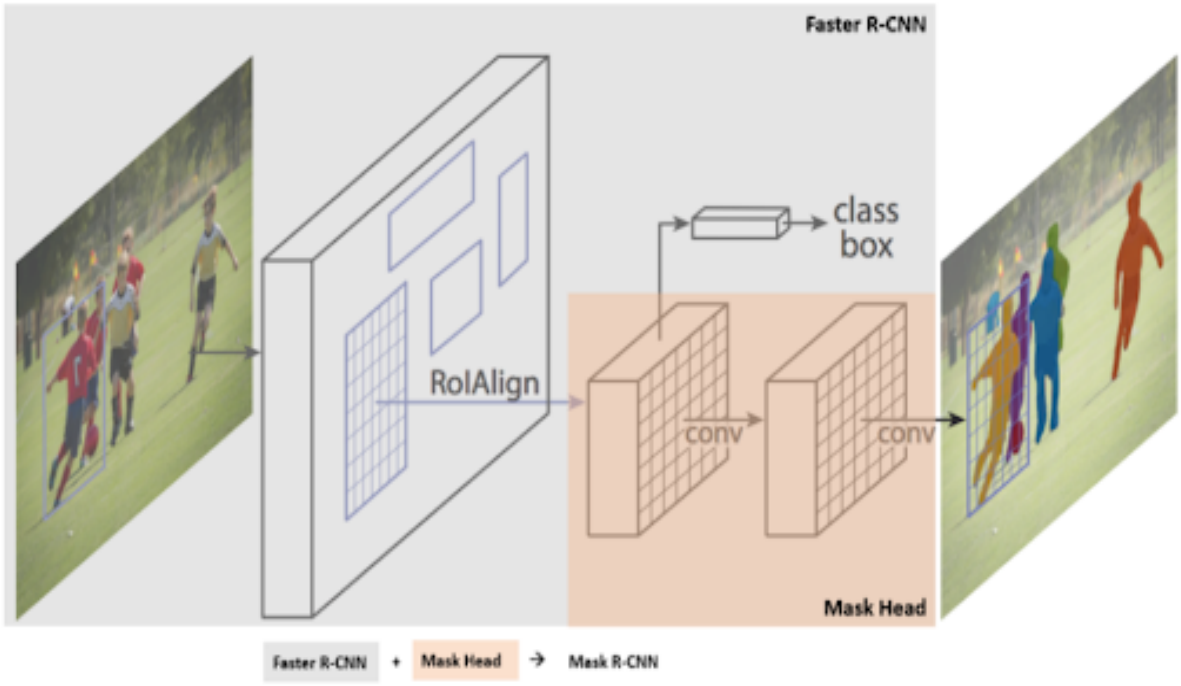


Figure 2.16: Mask RCNN [65]

The YOLO architecture network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. The convolutional layers are pretrained on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection, as shown in Figure 2.17.

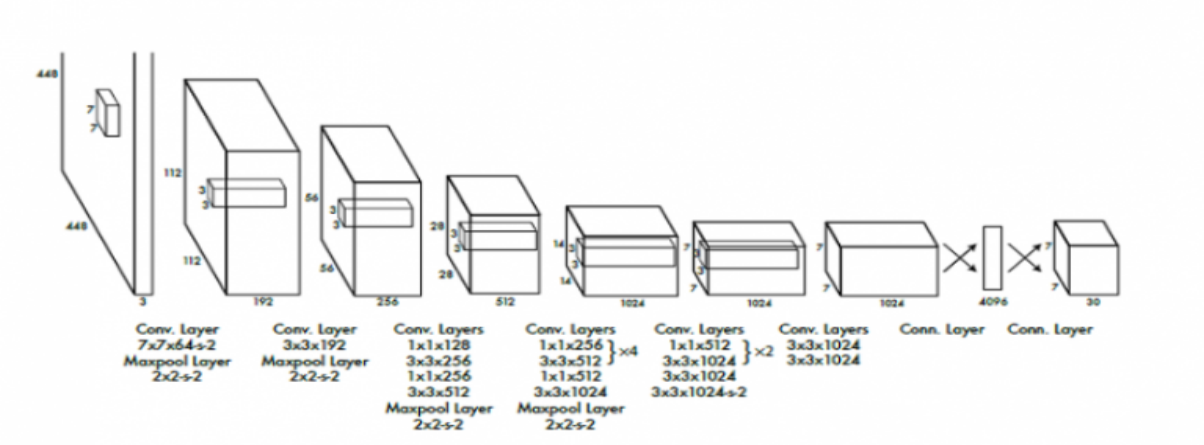


Figure 2.17: Yolo Architecture [66]

An SSD has two components: a backbone model and an SSD head. The backbone

model usually has a pre-trained image classification network as a feature extractor, typically a network like ResNet trained on ImageNet. The SSD head is just one or more convolutional layers added to this backbone, and the outputs are interpreted as the bounding boxes and classes of objects in the spatial location of the activations of the final layers, as shown in Figure 2.19.

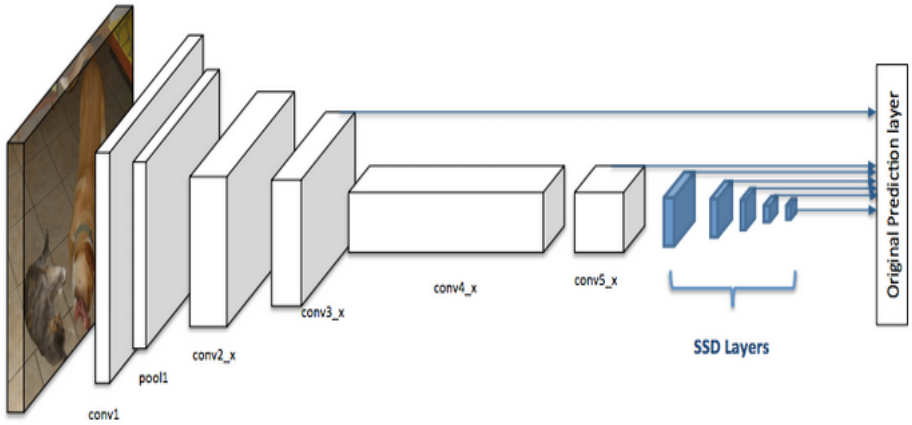


Figure 2.18: SSD Architecture [67]

2.4 Computer Vision

Some of the top computer vision techniques used with deep learning:

1. Image Classification
   Image classification refers to the task of identifying classes from an image.

2. Object Detection
   Object Detection refers to the task of identifying and locating objects in an image or video.

3. Semantic Segmentation
   Semantic Segmentation is the task of assigning a label to every pixel in the image.

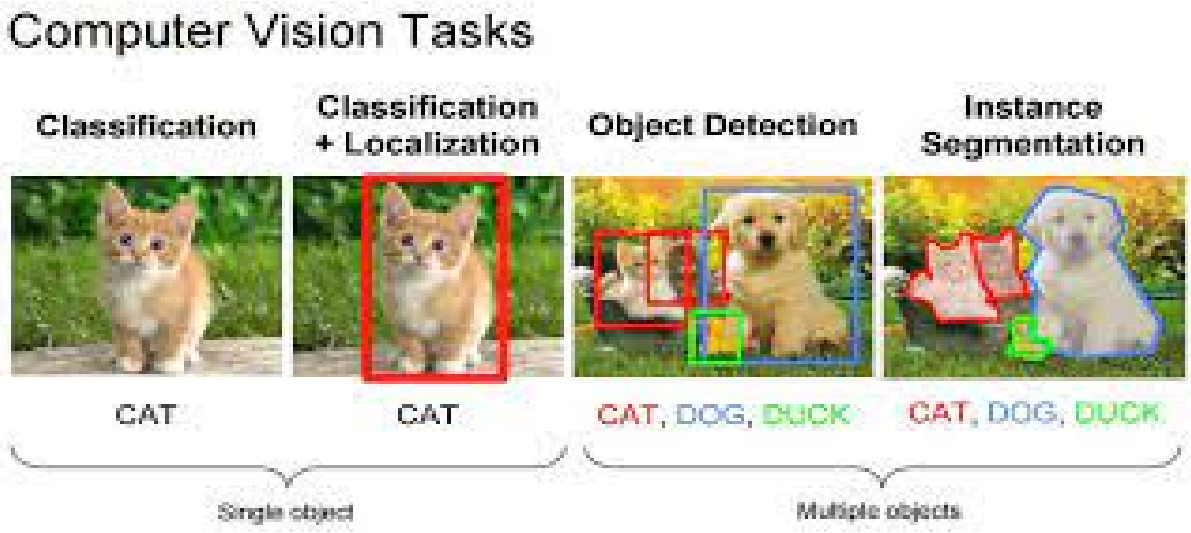


Figure 2.19: Computer vision techniques [68]

In addition, there are various image processing techniques that can be applied using openCV, an open-source computer vision library.

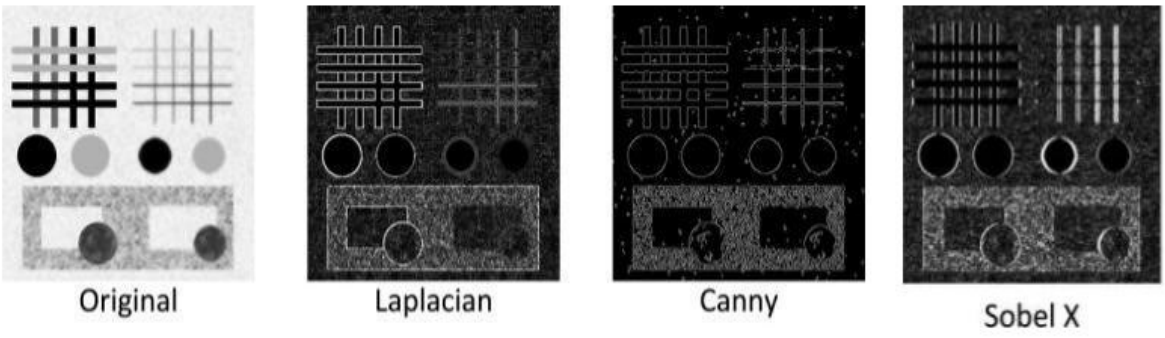


Figure 2.20: Representations of the different edge detection techniques [69]

There are two main types of edge detection [70]:

1. The gradient-based operator computes first-order derivations in a digital image. For Example, the Sobel operator and Prewitt operator.
2. The gaussian–based operator computes second-order derivations in a digital image. For example, Canny edge detector and Laplacian of Gaussian.

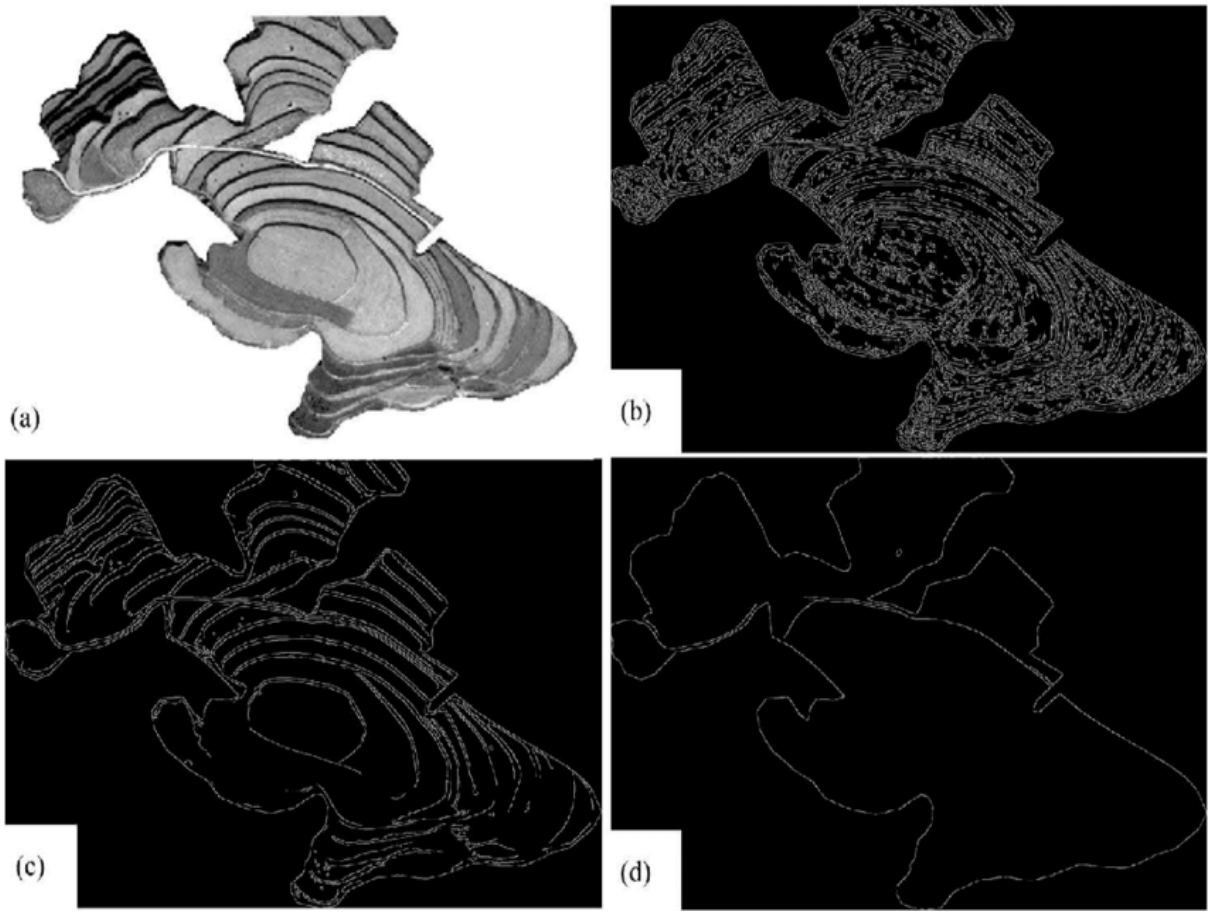


Figure 2.21: Representations of the different Canny edge thresholds techniques [69]

There are two main thresholds in Canny Edge detection [71]: upper and lower. If a pixel gradient is higher than the upper threshold, the pixel is accepted as an edge. If a pixel gradient value is below the lower threshold, then it is rejected. However, if the pixel gradient is between the two thresholds, then it will be accepted only if it is connected to a pixel that is above the upper threshold.

In Figure 2.21: (a) shows the terrace area; (b) shows the edge detection with upper and lower thresholds of 0.05 and 0.02, respectively, (c) has upper and lower thresholds of 0.2 and 0.08, respectively, and (d) has upper and lower thresholds of 0.5 and 0.2, respectively.

2.5 Literature Review

Indoor scene understanding presents a challenging recognition task due to the variety of backgrounds and the presence of rich and disordered decorative features.

Deep learning and CNNs have tremendously improved over the last few years and have shown great results in recognition tasks. Several studies have been conducted using deep learning techniques [72, 1, 73, 2] to tackle scene recognition for visual prostheses.

For general-purpose indoor scene recognition, a research paper [1] involved using many versions of efficientNet neural networks, fine-tuning them for indoor objects, and ranking their performances. The model was trained on the MIT 67 dataset and the indoor scenery was limited to 4 different categories (kitchen, bedroom, bathroom, and living room). The model was evaluated on MIT 67 and scene 15 datasets, achieving a high recognition rate, starting from 95%.

A previous paper published by Melina, et al. [72] used Mask-RCNN to detect objects in indoor scenes with some modifications to the classes. In addition, for the room layout, a solution proposed by Fernandez-Labrador et al. called PanoRoom, which involves the extraction of a 3D layout from a sphere was used. The selection of indoor scenes included bedrooms, bathrooms, kitchens, living rooms, dining rooms, and offices. The performance of the models used was tested using static images and videos. The tests were divided into two sections: object identification and indoor scene recognition. The correct responses for the recognition of room type using SIE-OM (Structural Informative Edges and Object Masks) (54%) were higher than the OM method (46%) for static images, implying that structural edges improve recognition of indoor scenes. Also, the experiments with videos showed an increase in the percentage of correct responses for both SIE-OM and OM methods.

Another paper by the same publisher used a different approach for structural edge detection suggested by Mallya, et al. [1] and the same approach for object detection (Mask-RCNN). Three different methods were compared: Canny, OM, and SIE-OM with SIE-OM showing the best results (55%), and Canny showing the least recognition rate (32%). SIE-OM was tested using 10 different indoor images. Subjects were asked to identify the room type and the level of certainty. Overall, the total percentage of correct object identification tasks was very high, at 88% and the room type recognition was 55% on average.

Further work has been done in another paper [2] following the same approach to the object mask extraction section where refinements such as sorting object masks by probability scores to avoid object occlusions, as well as adding changing the colour of object masks to gray and leaving the object silhouettes (outline of the object) white after applying the Mask-RCNN model to enhance object recognition have been made. The SIE-OMS method has the highest percentage of correctly identified objects (62.78%) compared to Edge (19.17%) and Direct (36.83%) methods. Additionally, the SIE-OMS method has the highest percentage of correctly identified room types (70.33%) compared to Edge (13.33%) and Direct (35.33%) methods.

In all the aforementioned papers, the prosthetic vision was simulated using a Gaussian Luminance profile based on previous studies on simulated prosthetic vision [46].

Recent research has developed a framework for simulating prosthetic vision called pulse2percept [74]. Various types of retinal implants, including ArgusII, AlphaAMS, and AlphaIMS, are all accommodated in the library. Furthermore, computational models for focal percepts (scoreboard model) and axonal streaks (axon map model) have been created and are available for use, allowing us to evaluate the vision of retinal prosthesis patients more easily and accurately. A series of linear filtering and nonlinear processing steps that model the spatial and temporal sensitivity of the retinal tissue process the electrical stimulus. In addition, patients were asked to report the intensity, brightness, or size of percepts on a rating scale. The output of the model is a prediction of the visual perception, which can be compared to patients' drawings and descriptions of what they see.

Chapter 3

Methodology

3.1 Approach Overview

First, the image is obtained, and then image processing techniques are applied to it. After processing the image, it is resized to 32 by 32 pixels and a phosphene simulation is obtained.

There are three main image processing techniques (enhancement, edges, and direct). The enhancement group is where Mask-RCNN is used and image processing techniques such as contrast and padding are enhanced, as shown in Figure 3.1. The Edges group makes use of the Canny edge detection algorithm, as shown in Figure 3.3, and the Direct group is where no image processing technique is applied, as shown in Figure 3.2.

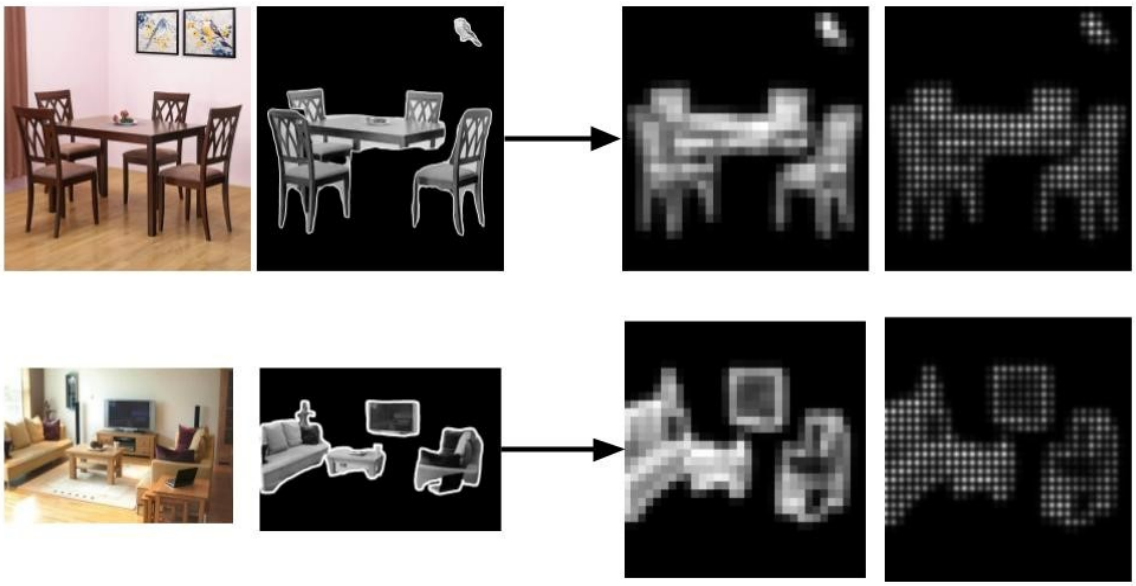


Figure 3.1: Enhancement approach

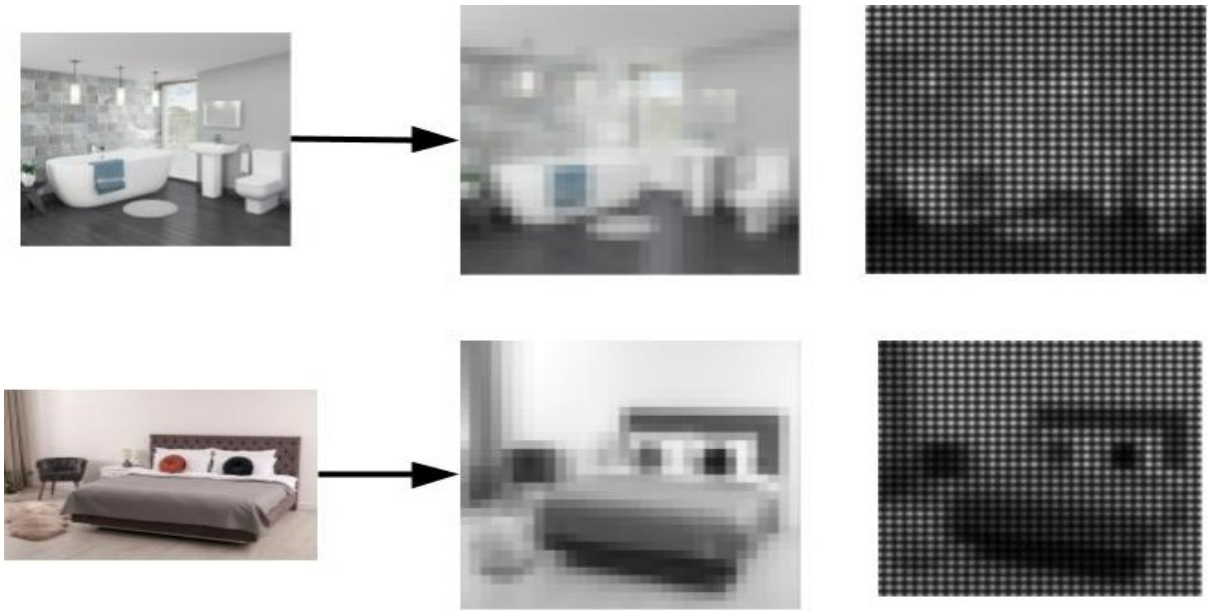


Figure 3.2: Direct approach

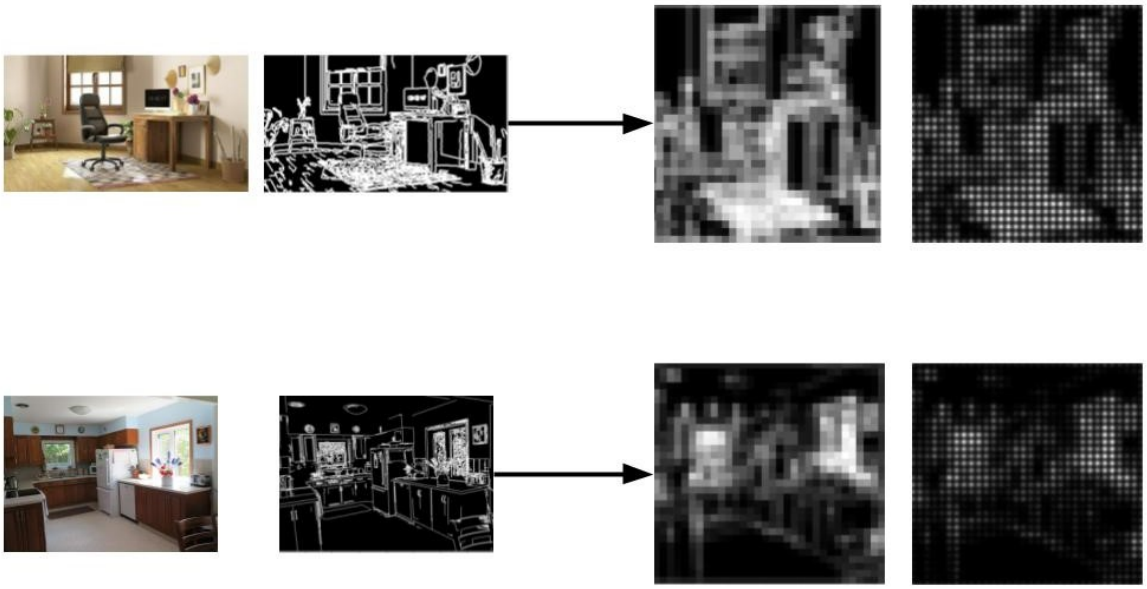


Figure 3.3: Edges approach

3.2 Methods

3.2.1 Edge detection

The most commonly used edge detection algorithms in OpenCV are Canny and Sobel, both of which belong to separate edge detection types. Both types have their own benefits and limitations. Sobel is known for its simple and time-efficient computation. However, it is highly sensitive to noise and not very accurate in edge detection. Canny, on the other hand, is less sensitive to noise and has good localization without altering the features, which means better accuracy in edge detection.

Despite the fact that Sobel is faster, we were mainly working on single images, so time was not a limiting factor and accuracy mattered more. Accordingly, we have settled on Canny edge detection.

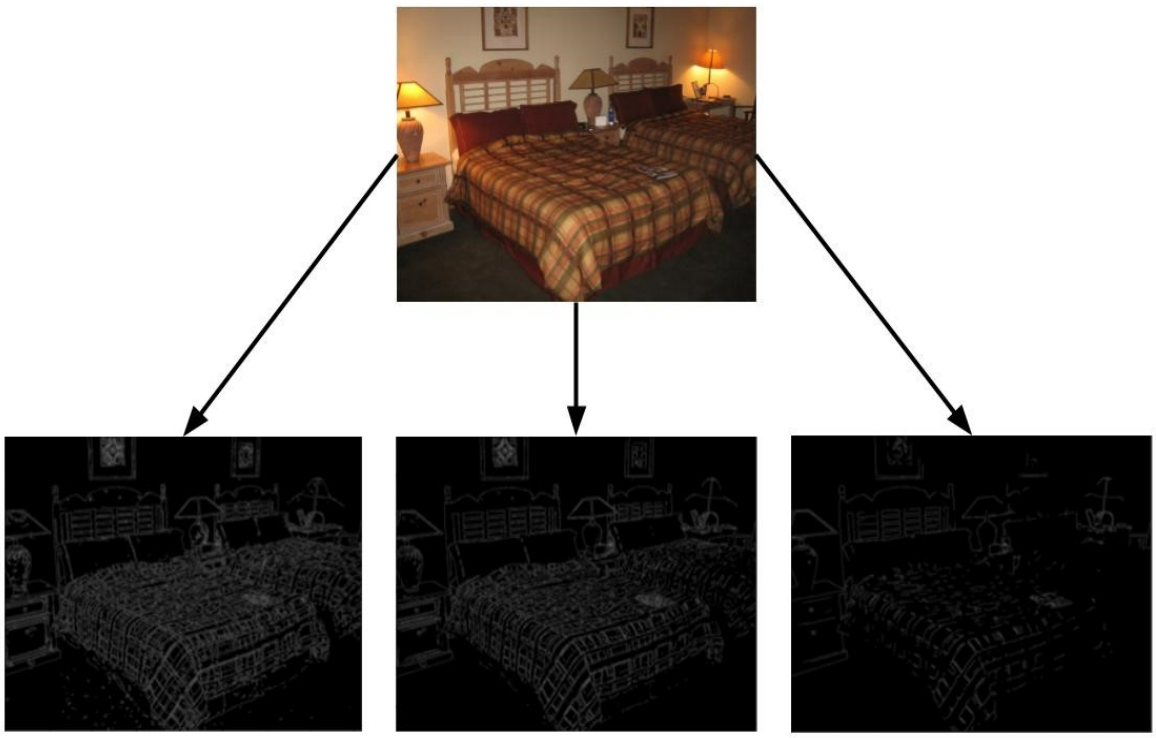


Figure 3.4: Representations of the effect of different canny thresholds.

In Figure 3.4, the first output image has a lower threshold of 20 and an upper threshold of 50. The second output image has a lower threshold of 50 and an upper threshold of 80. The third output image has a lower threshold of 80 and an upper threshold of 120.

In the implemented experiment, it was decided that the best values for the upper and lower thresholds were 50 and 80, respectively, because they reserved the most important edges with the minimum possible noise.

Further processing was needed to make the edges clearer. Thus, the morphological operator (dilation) was applied to make the lines thicker and easier to see, as shown in Figure 3.5.

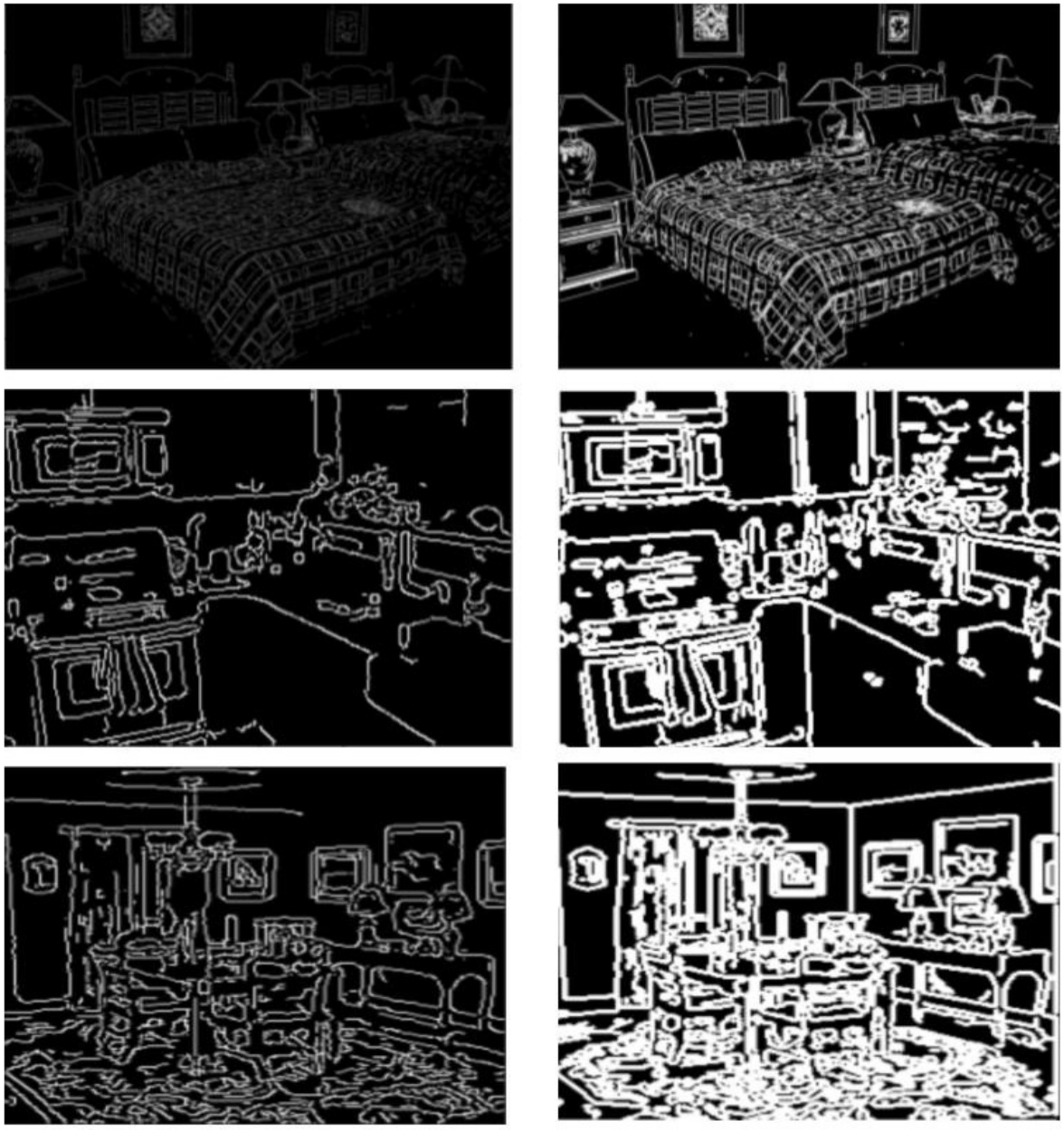


Figure 3.5: Left column shows images before dilation and right column shows images after edges dilation

3.2.2 Object Segmentation

In this research, we used Mask R-CNN because it allows for applying image processing techniques more conveniently since we can have access to the bounding boxes as well as the masks. Mask R-CNN is also simple to train, and it outperforms all existing, single-model entries on every task. Furthermore, the method is very efficient and adds only a small overhead to Faster R-CNN.

The Mask-R CNN model used is pre-trained on the COCO dataset. The MS COCO dataset is a large dataset developed by Microsoft specifically for object detection and segmentation. This method was developed using Python and the TensorFlow framework. There are 91 object categories, which include individual instances that may be easily recognized (person, bed, chair, etc.). Since we are only interested in indoor scenes, some of the objects we need include a hat, backpack, umbrella, shoe, eyeglasses, handbag, tie, suitcase, bottle, plate, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted plant, bed, mirror, dining table, desk, toilet, door, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, blender, book, clock, vase, scissors, teddy bear, hair drier, toothbrush, and hairbrush. All of these objects are likely to be present in indoor scenes like living rooms, bathrooms, bedrooms, kitchens, dining rooms, and offices.

One limitation of Mask-RCNN is that it does not identify the same object within separate frames, resulting in assigning different colour masks to the same object within successive frames in a video, as shown in Figure 3.6.

Figure 3.6: Colours of the same object change within the same video.

First, the class of the object is identified. Then, in subsequent frames, we check if the same object class was present in the previous frames. If so, IOU (Intersection Over Union) is applied to check if this object is indeed the same as the object in the previous

$$IOU = (AreaOfOverlap)/(AreaOfUnion)$$

Figure 3.7: Intersection Over Union (IOU)

frames. For an average-paced video speed, a threshold of 0.2 does the job. Intersection over union is an evaluation metric used to evaluate how accurate the model is at predicting the object's location, as shown in Figure 3.7.

Initially, binary masks were extracted from segmentation and then converted to phosphenes. However, this leads to the loss of significant object details, as shown in Figure 3.8.

Figure 3.8: Binary masks extraction from Mask-RCNN

The gray-scale version was used instead of the binary version, adding more details to the image. Enhancement techniques, including thicker white padding around the MASK-RCNN masks and increasing the image contrast using a histogram equalizer, were used, as shown in Figures [3.9, 3.10]).

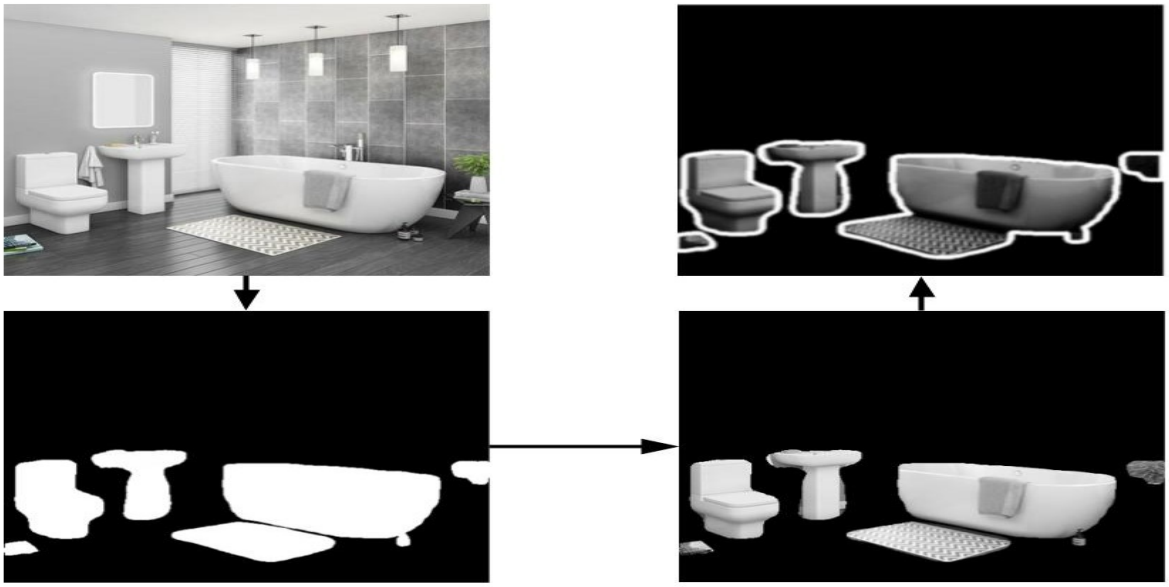


Figure 3.9: Final Output of Mask-RCNN

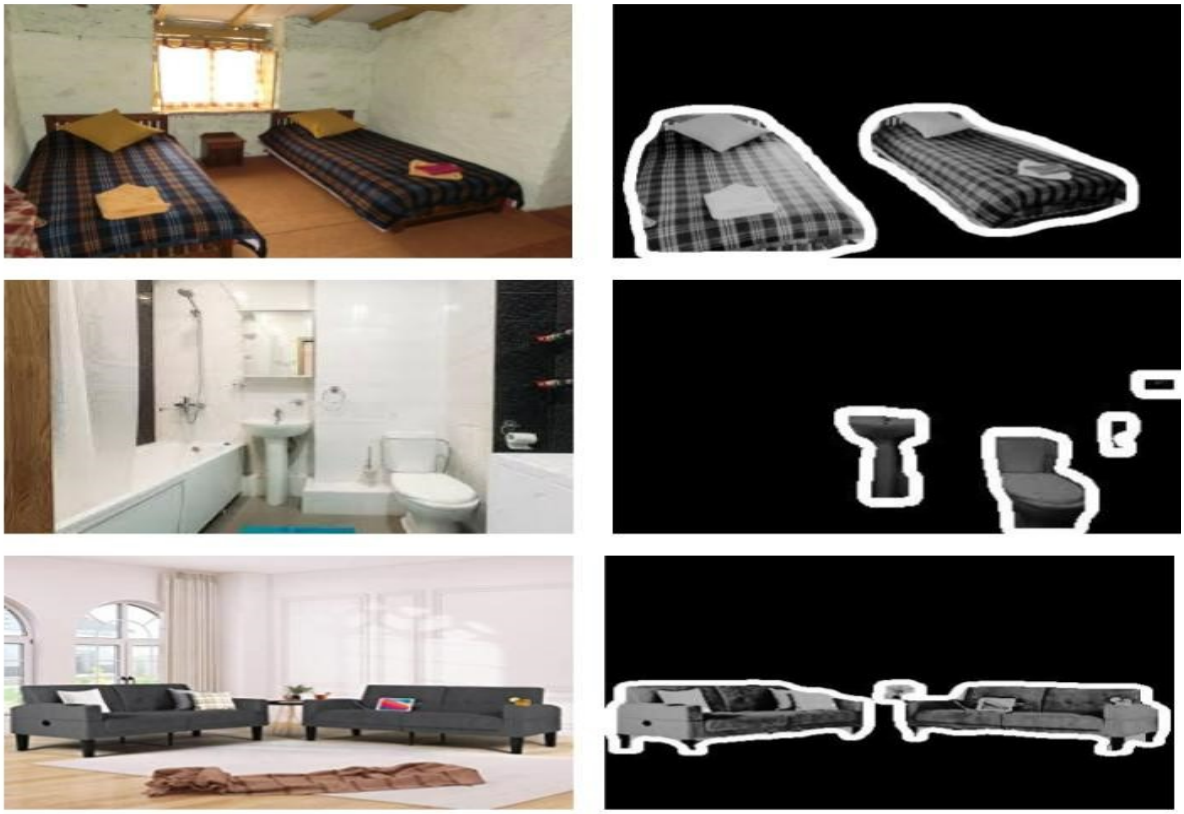


Figure 3.10: More examples of the final output obtained from Mask-RCNN

3.3 Phosphene simulation

For creating the phosphene simulation, a library, Pulse2percept [74], is used. Pulse2percept is an open-source Python simulation framework used to predict the perceptual experience of retinal prosthesis patients.

In visual prostheses, there is a wide range of implant configurations. We will follow a configuration similar to the previous implementation in the recent visual prosthesis-related papers using the scoreboard model described in [Beyeler2019], where all percepts are Gaussian blobs and the implant used is the Argus II Retinal Prosthesis System.

Since 32 x 32 images were the main concern, the Argus II implant grid size was also set to 32 x 32 and the visual field angle was also set to 32° to convert the whole input image into phosphenes. The visual field will later be adjusted while performing the experiment.

The phosphene size (rho) was set to 70 microns and the spacing was adjusted to accommodate all the phosphenes within the 32 by 32 grid, as shown in Figure 3.11.

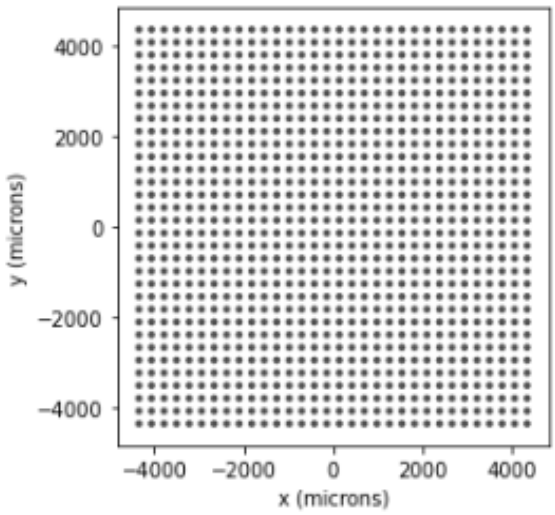


Figure 3.11: Argus II implant used in the implementation

The scoreboard model was used, producing rounded phosphenes at each corresponding electrode, with brightness levels that depends on the light intensity, as shown in Figure 3.12.

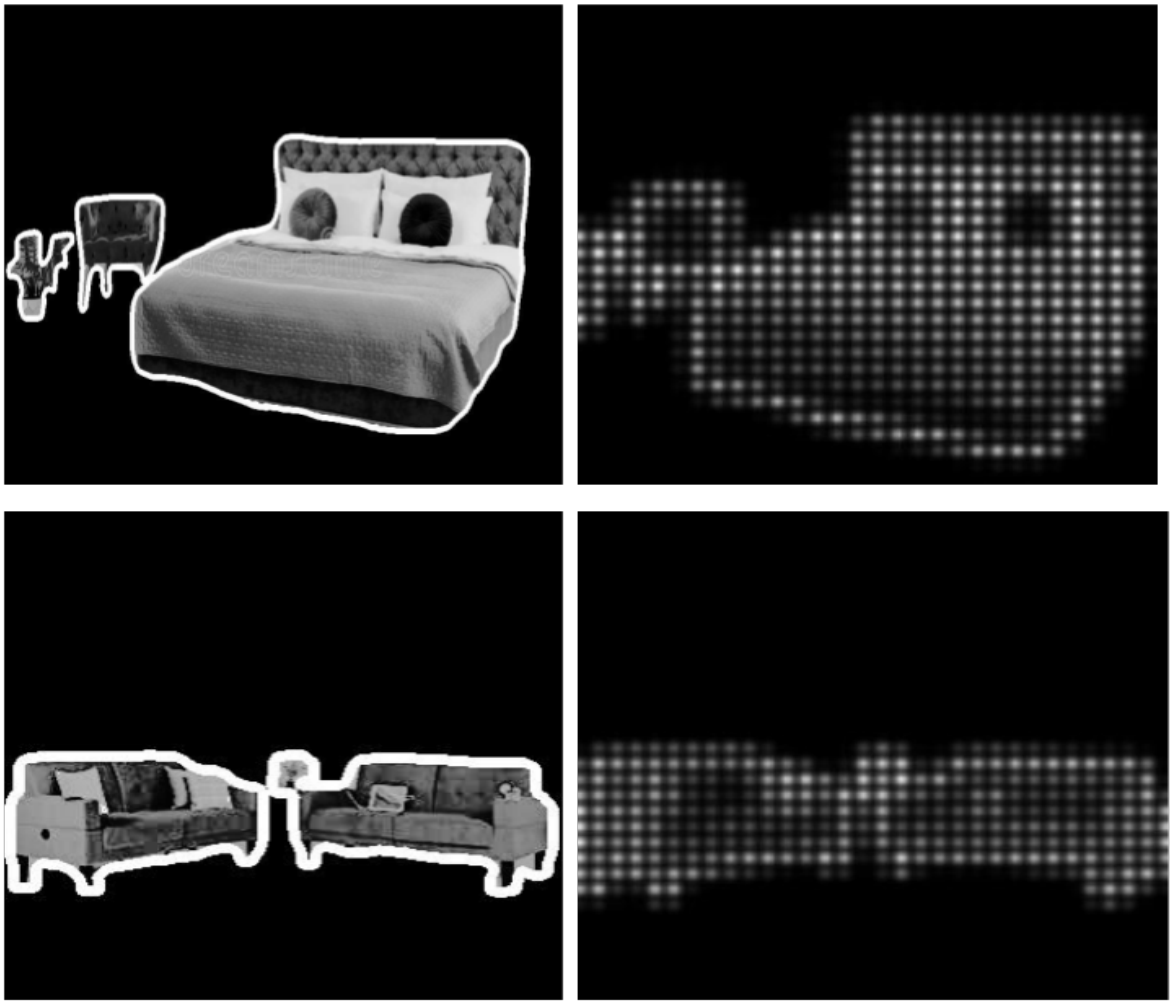


Figure 3.12: More examples of the final output obtained from phosphene simulation

3.4 Experiments setup and Methodology

There are three separate experiments in this research: screen, VR, and real-time.

The experiments were non-invasive. Participants were informed of the experiment and their consent was taken before conducting the experiments on them. All the participants were volunteers, and they could leave the experiment at any time.

3.4.1 Experiment 1: Screen

The experiment involved static images and videos displaying several different indoor scenes. There were 18 participants, 6 for each group, ages between 14 and 22, and they included both males and females. The mean age is 20.22 years old with a standard deviation of 1.76. Participants were asked to identify room types and the objects within them.

Six main indoor scenes were used: bedroom, bathroom, office, dining room, kitchen, and living room. Participants were placed at a distance of about 1 m to maintain a visual field of 20°, as shown in Figure 3.13.

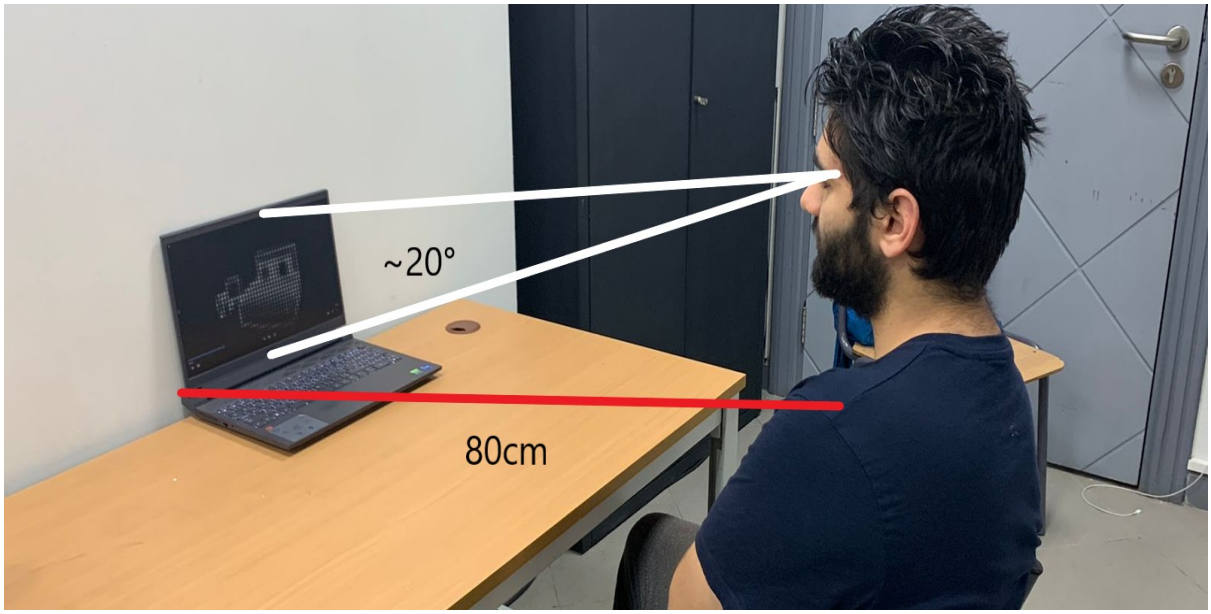


Figure 3.13: Screen Experiment

Three groups were tested:

1. Enhancement group
   This involved the output of MASK-RCNN and image processing.

2. Direct group
   No enhancement technique was applied, i.e, no image processing was applied.

3. Edges group
   This involved extracting the edges from the original images.

3.4.2 Experiment 2: VR

STEAMVR and the HTC Vive headset were used to simulate prosthetic vision in virtual reality, as shown in Figure 3.14. A different set of images were used in the VR experiment. Each group was composed of six participants, ages 19 to 22. The mean age is 20.58 years with a standard deviation of 0.793.

Figure 3.14: VR Experiment

Two groups were tested:

1. Enhancement group: This involved the output of MASK-RCNN and image processing.
2. Direct group: No enhancement technique was applied, i.e, no image processing was applied.

3.4.3 Experiment 3: Real-time

The experiment was repeated but in real time. Participants wore a VR Box headset and the phone camera was attached to the outside of the headset to capture the video, as shown in Figure 3.15.

Only computer rooms were used in the experiment because they were available, as shown in Figure 3.16. Objects were brought into the room and participants were asked to identify the room, navigate through the room and identify the objects within the room, as shown in Figure 3.17.

Objects that were present in the room: computer screens, keyboards, laptops, tables, chairs, a potted plant, a remote controller, a stopwatch, and a black box.

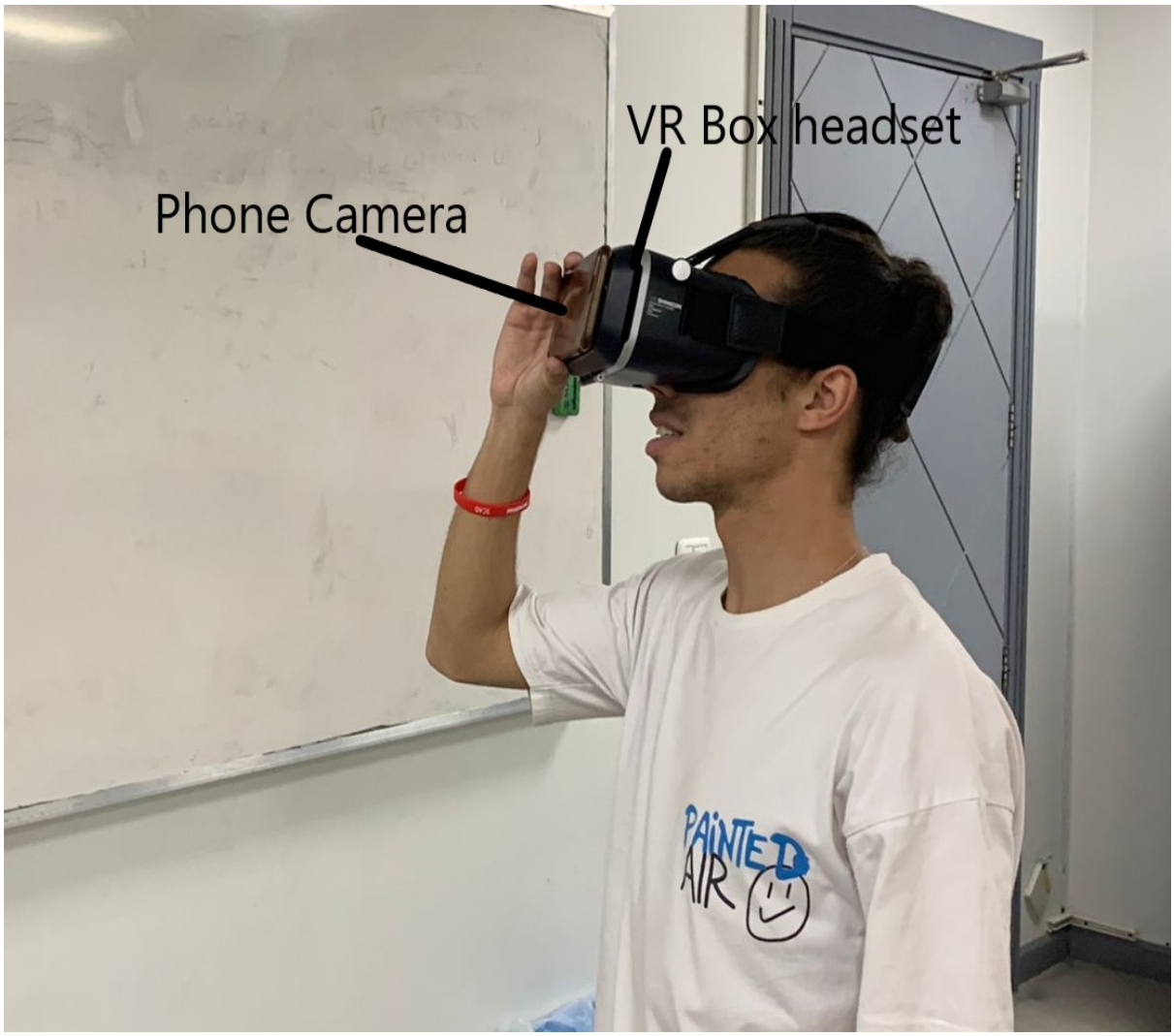


Figure 3.15: Participant wearing the VR Box headset to simulate real time prosthetic vision.

Two groups were tested:

1. Enhancement group
   This involved the output of MASK-RCNN and image processing.

2. Direct group
   No enhancement technique was applied, i.e, no image processing was applied.

Each group had three participants, ages ranging from 21 to 22. The average age is 21.17, with a standard deviation of 0.408.

Figure 3.16: Lab where the experiment was conducted.

Figure 3.17: Some examples of the objects the participant was asked to identify in the room.

Chapter 4

Results

4.1 Pre-processing results

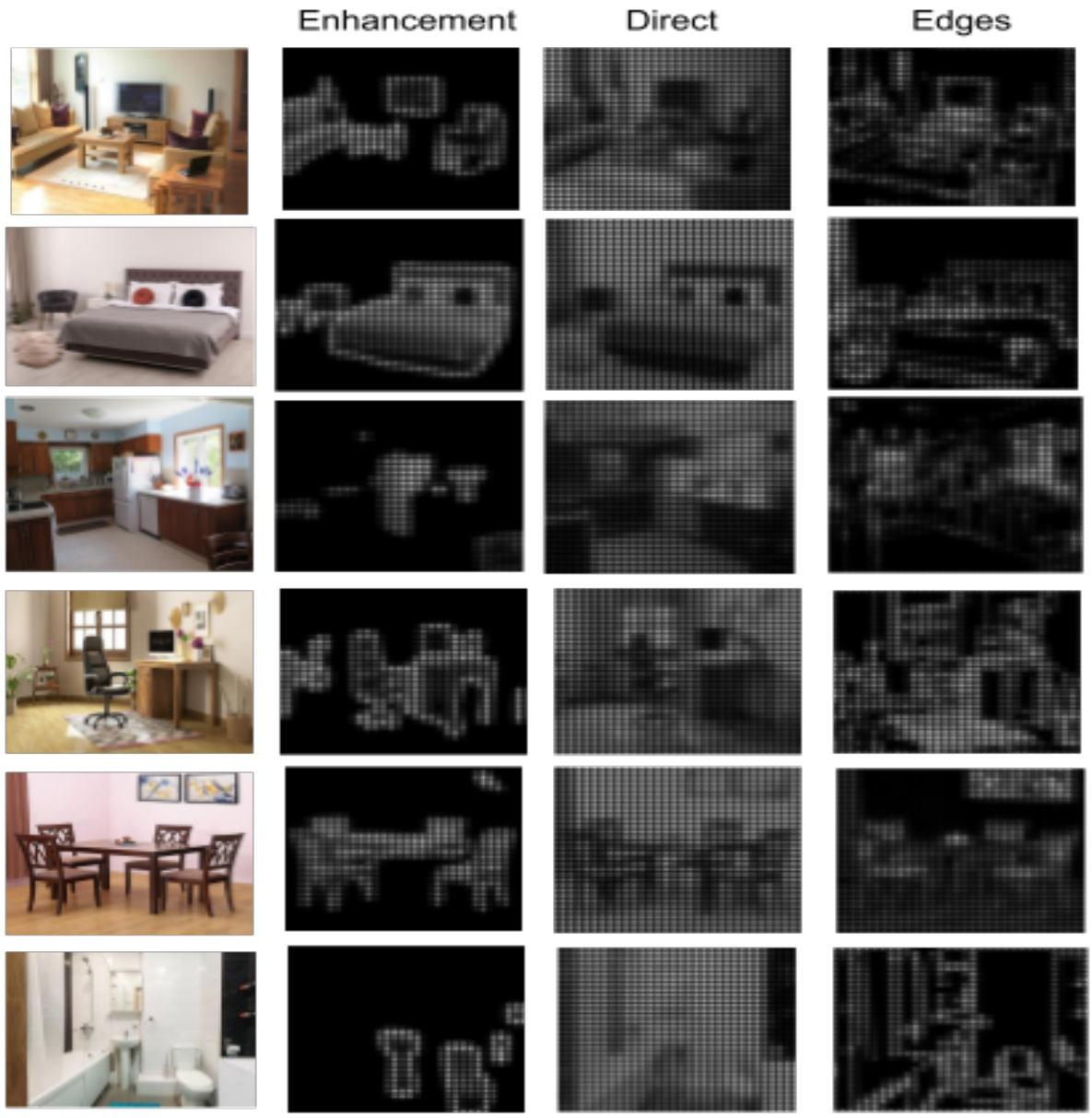


Figure 4.1: Examples of stimuli used in the experiment.

Six examples of indoor environments represented with 1024 phosphenes (rows: living room, bedroom, dining room, kitchen, office and bathroom, respectively). Each column in Figure 4.1 shows: a) input images, b) images processed using the Enhancement method, c) images processed using the Direct method and d) images processed by Edges method, respectively.

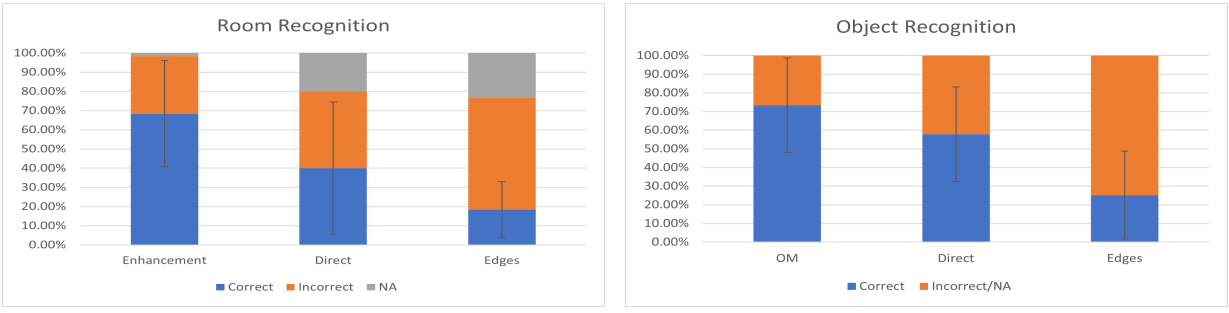
4.2 Experiment 1 (Screen) results

4.2.1 Image stimuli results

Table 4.1: Comparison of responses of the three different methods (Enhancement , Direct and Edges) on object identification and room type recognition tasks using images.

Method	% Room recognized	% Object Identification	NA	1	2	3	4	5
Enhancement	68.33%	73.33%	1.67%	6.67%	18.33%	20.00%	18.33%	35.00%
Direct	40.00%	57.78%	20.00%	16.67%	21.67%	10.00%	11.67%	20.00%
Edges	18.33%	25.17%	23.33%	5.00%	6.67%	13.33%	10.00%	8.33%

Table 4.1 shows the overall responses to the three different methods. The assurance for room identification of the three methods was evaluated on a scale of 1 to 5, with 5 being the highest level of confidence. It is clear that participants in the enhancement are generally more confident in their responses than in the Direct and Edges group. According to Table 4.1, the enhancement method had a higher percentage of correct responses for the room type recognition (68.33%) than the Direct (40%) and the Edges method (18.33%).The results for each group conclude that there is a significant difference between the three groups ($p<0.001$), according to the ANOVA test. Furthermore, there is a significant difference between enhancement and direct ($p<0.05$) as well as enhancement and edges ($p<0.05$). Edges show a greater statistical difference between the other two methods. Furthermore, the percentage of participants with a higher level of confidence (4-5) in the Enhancement group (OM) was higher than for the Direct and Edges groups.



(a) (b)

Figure 4.2: Overall results for room identification and object recognition task for images. Percentage of correct, incorrect and not answered responses.

Higher scores in correct responses indicate that subjects were able to identify and recognize the objects and the type of room in each test image. Higher ratios of not answered indicate that subjects were not able to identify and recognize the objects and the type of room in each test image. According to the results displayed on Fig.4.2, the general findings are that: Enhancement method improves the identification of the objects resulting to be the most effective method.

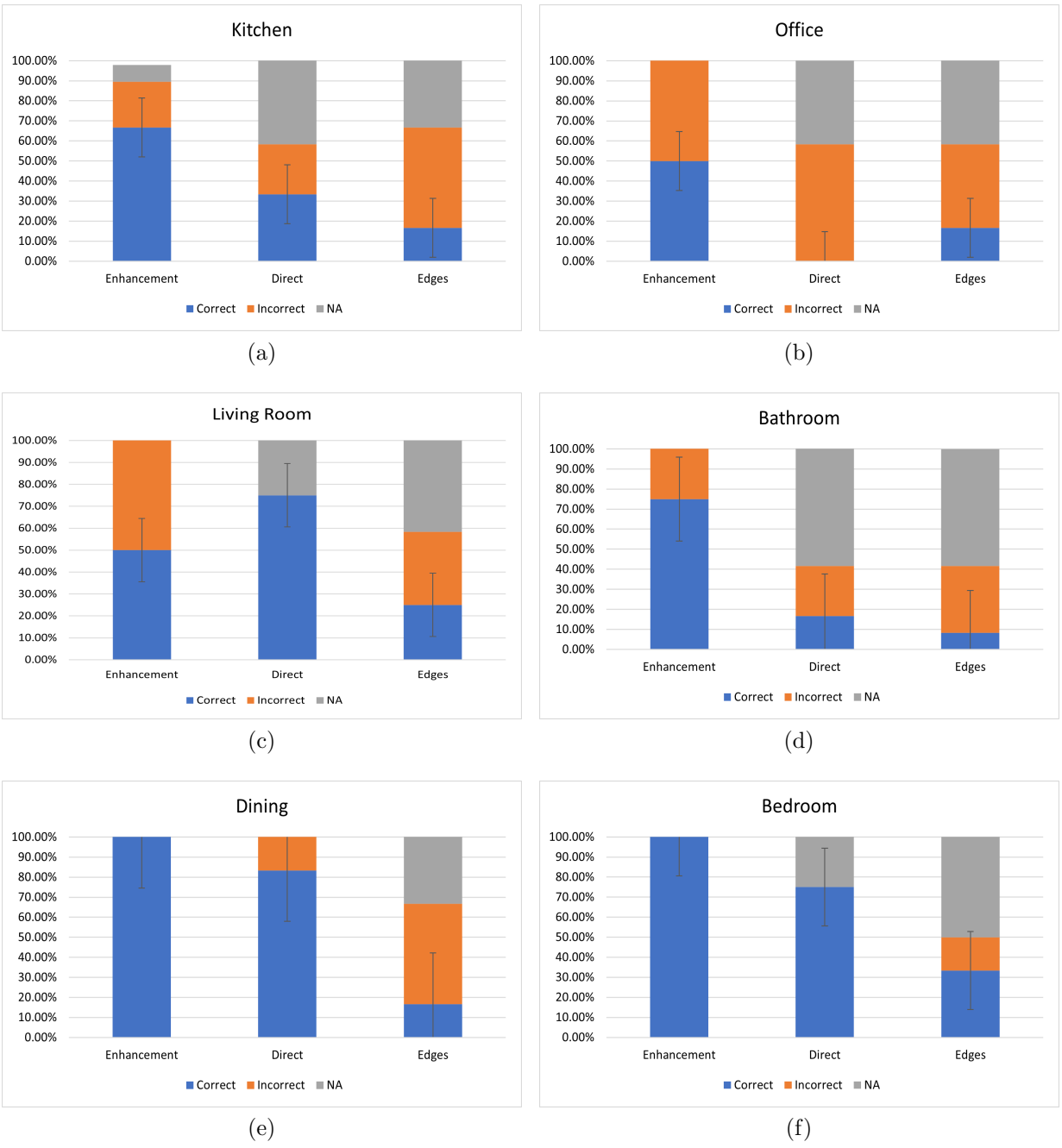


Figure 4.3: Room identification results for each room-type.

Higher scores in correct responses indicate that subjects were able to identify and recognize the objects and the type of room in each test image. Higher ratios of not answered indicate that subjects were not able to identify and recognize the objects and the type of room in each test image. According to the results displayed in Figure 4.2, the general findings are that the enhancement method improves the identification of the objects, resulting in it being the most effective method.

Table 4.2: Confusion matrix results of room type recognition by using the Enhancement method.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.75	0.08	0.00	0.00	0.08	0.08	0.00	1.00	75.00%
Kitchen	0.17	0.67	0.08	0.00	0.00	0.00	0.08	1.00	66.67%
Bedroom	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	100.00%
Living Room	0.17	0.00	0.08	0.58	0.17	0.00	0.00	1.00	58.33%
Office	0.00	0.00	0.08	0.00	0.50	0.42	0.00	1.00	50.00%
Dining Room	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	100.00%
Total	1.08	0.75	1.25	0.58	0.75	1.50	0.08	6.00	
Precision	69.23%	88.89%	80.00%	100.00%	66.67%	66.67%	0.79		

The results in Table 4.2 show that some participants confused room types, perhaps due to the similarity of the objects present in them. For example, a larger percentage of participants confused the office with the dining room because of the presence of tables and chairs. Some participants confused the living room with the bedroom, mistakenly identifying the couch as a bed.

Table 4.3: Confusion matrix results of room type recognition by using the Direct method.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.75	0.08	0.00	0.00	0.08	0.08	0.00	1.00	75.00%
Kitchen	0.00	0.33	0.00	0.25	0.00	0.00	0.42	1.00	33.33%
Bedroom	0.00	0.00	0.67	0.17	0.00	0.17	0.00	1.00	66.67%
Living Room	0.00	0.00	0.00	0.75	0.00	0.00	0.25	1.00	75.00%
Office	0.25	0.08	0.08	0.08	0.00	0.08	0.42	0.92	0.00%
Dining Room	0.00	0.00	0.00	0.00	0.17	0.83	0.00	1.00	83.33%
Total	1.00	0.50	0.75	1.25	0.25	1.17	1.08	5.92	
Precision	75.00%	66.67%	88.89%	60.00%	0.00%	71.43%			

Table 4.4: Confusion matrix results of room type recognition by using the Edges method.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.75	0.08	0.00	0.00	0.08	0.08	0.00	1.00	75.00%
Kitchen	0.17	0.17	0.08	0.08	0.00	0.08	0.33	0.92	18.18%
Bedroom	0.00	0.00	0.33	0.00	0.00	0.17	0.50	1.00	33.33%
Living Room	0.00	0.17	0.17	0.25	0.00	0.00	0.42	0.83	25.00%
Office	0.00	0.08	0.08	0.08	0.17	0.17	0.42	0.92	16.67%
Dining Room	0.17	0.00	0.00	0.17	0.17	0.17	0.33	1.00	16.67%
Total	1.08	0.50	0.67	0.58	0.42	0.67	2.00	5.67	
Precision	69.23%	33.33%	50.00%	42.86%	40.00%	25.00%			

Overall, as derived from the above confusion matrices (Tables[4.2, 4.3, 4.4]) , edges tend to have low precision values (generally below 50%, meaning that rooms were identified correctly with a probability of less than 50%). Direct and enhancement groups tend to have higher precision values, with the average precision value for the enhancement group being around 80% while for the direct method it's more toward 60%.

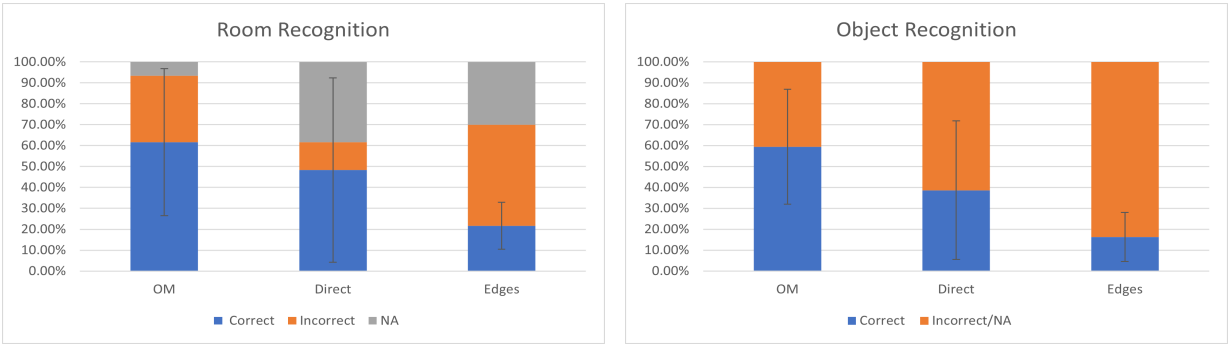
4.2.2 Video stimuli results

The same experiment was repeated using videos as input stimuli instead of static images. Again, the enhancement technique of obtaining the object masks (OM) showed the best results. The assurance for room identification of the three methods was evaluated on a scale of 1 to 5, with 5 being the highest level of confidence.

Table 4.5: Comparison of responses of the three different methods (Enhancement , Direct and Edges) on object identification and room type recognition tasks.

Method	% Room recognized	% Object Identification	NA	1	2	3	4	5	Recall
OM	61.67%	59.44%	6.67%	8.33%	16.67%	21.67%	16.67%	21.67%	75.00%
Direct	48.33%	38.61%	38.33%	6.67%	10.00%	6.67%	20.00%	18.33%	18.18%
Edges	16.67%	16.33%	30.00%	5.00%	16.67%	10.00%	13.33%	8.33%	33.33%

Again, the enhancement technique of obtaining the object masks (OM) showed the highest % of room recognition. However, this time, the difference between enhancement (OM) and the direct method is smaller compared to using images.



(a) (b)

Figure 4.4: Overall results for room identification and object recognition task for videos.Percentage of correct, incorrect and not answered responses.

Higher scores in correct responses indicate that subjects were able to identify and recognize the objects and the type of room in each test image. Higher ratios of not answered indicate that subjects were not able to identify and recognize the objects and the type of room in each test image. The overall accuracy of the enhancement technique decreased, while for the direct as well as the edge methods, the percentage of correct responses decreased (Fig 4.4). One explanation for this was that it was due to the flickering of the video that uses the enhancement technique, causing the participant to

be confused. One way to tackle this in the future to obtain more accurate results is to move the camera extremely slowly while taking the video. Meanwhile, one explanation for the increased percentage of correct responses for the direct and edge methods is the changing light intensity in the video, which allows the participant to make sense of the objects being displayed and identify their outer frames.

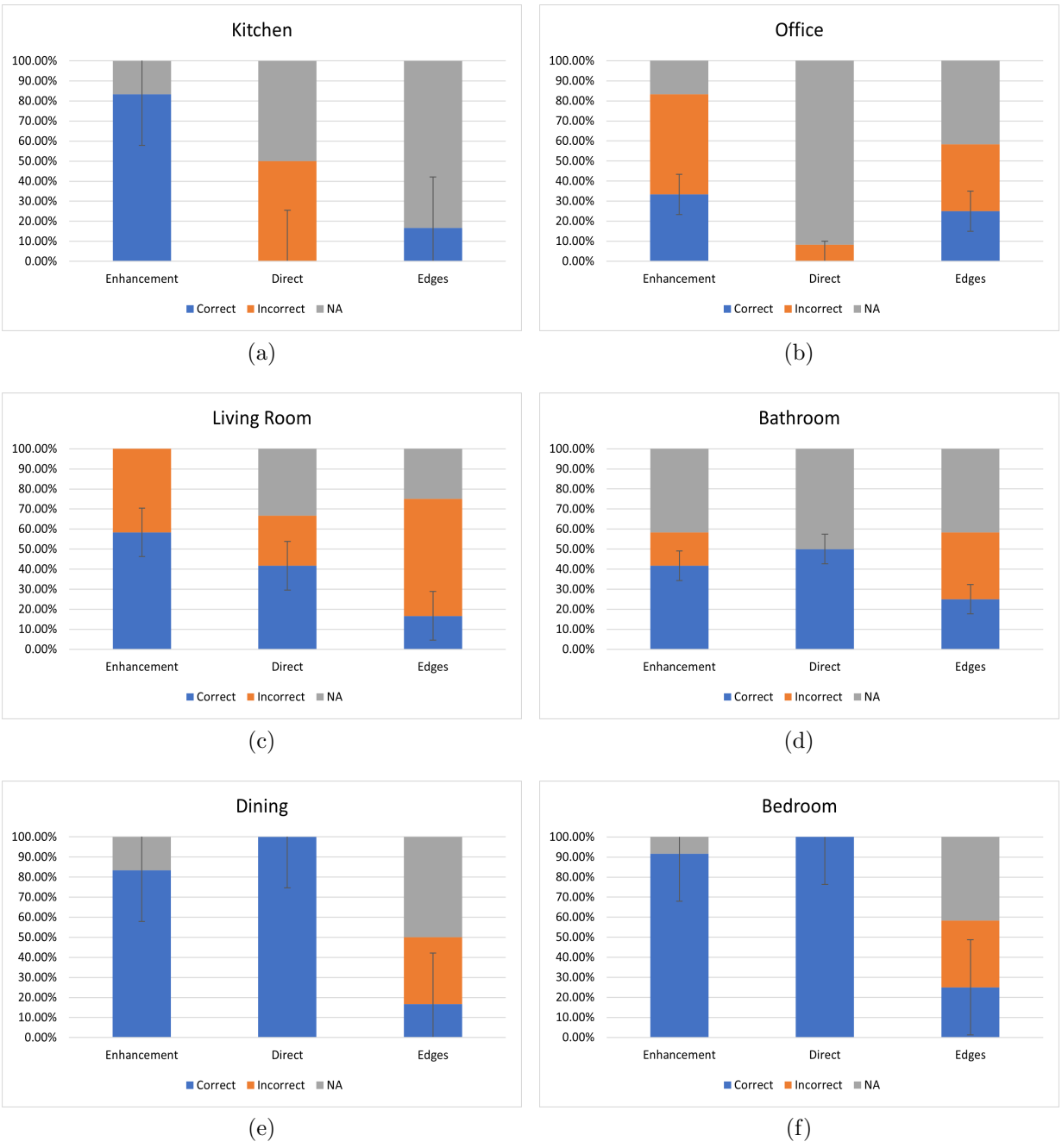


Figure 4.5: Room identification results for each room-type.

Room identification results for each room type. Higher scores in correct responses indicate that subjects were able to recognize the type of room in each test image. Higher ratios of non-responses indicate that subjects were not able to recognize the type of room in each image. According to the results displayed on Figure 4.5, the Enhancement and Direct methods usually obtain the highest score for room type identification compared with the Edge and Direct methods.

Table 4.6: Confusion matrix results of room type recognition by using the Enhancement method.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.42	0.00	0.00	0.00	0.17	0.00	0.42	1.00	41.67%
Kitchen	0.00	0.83	0.00	0.00	0.00	0.00	0.17	1.00	83.33%
Bedroom	0.00	0.00	0.92	0.00	0.00	0.00	0.08	1.00	91.67%
Living Room	0.08	0.08	0.17	0.58	0.08	0.00	0.00	1.00	58.33%
Office	0.00	0.08	0.00	0.00	0.33	0.42	0.17	1.00	33.33%
Dining Room	0.00	0.00	0.00	0.00	0.00	0.83	0.17	1.00	83.33%
Total	0.50	1.00	1.08	0.58	0.58	1.25	1.00		
Precision	83.33%	83.33%	84.62%	100.00%	57.14%	66.67%			

Table 4.6 shows that some participants in the enhancement group confused room types, perhaps due to the similarity of the objects present in them. The most noticeable confusion was between the office and the dining room. A large percentage of participants were unable to identify the bathroom correctly.

Table 4.7: Confusion matrix results of room type recognition by using the Direct method.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.50	0.00	0.00	0.00	0.00	0.00	0.50	1.00	50.00%
Kitchen	0.33	0.00	0.00	0.17	0.00	0.00	0.50	1.00	0.00%
Bedroom	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	100.00%
Living Room	0.00	0.00	0.25	0.42	0.00	0.00	0.33	1.00	41.67%
Office	0.00	0.00	0.00	0.08	0.00	0.00	0.92	1.00	0.00%
Dining Room	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	100.00%
Total	0.83	0.00	1.25	0.67	0.00	1.00	2.25		
Precision	60.00%	0.00%	80.00%	62.50%	0.00%	100.00%			

Table 4.7 shows that all participants in the direct group were able to identify the bedroom and the dining room correctly. None were able to identify the office, and a large percentage of participants confused the bedroom with the living room, more likely because they identified the couch as a bed.

Table 4.8: Confusion matrix results of room type recognition by using the Edges method for videos.

Actual/Predicted	Bathroom	Kitchen	Bedroom	Living Room	Office	Dining Room	NA	Total	Recall
Bathroom	0.17	0.08	0.08	0.17	0.00	0.17	0.33	1.00	16.67%
Kitchen	0.00	0.17	0.00	0.00	0.00	0.00	0.83	1.00	16.67%
Bedroom	0.00	0.08	0.25	0.08	0.17	0.00	0.42	1.00	25.00%
Living Room	0.00	0.08	0.33	0.17	0.00	0.17	0.25	1.00	16.67%
Office	0.00	0.00	0.00	0.17	0.08	0.17	0.58	1.00	8.33%
Dining Room	0.00	0.00	0.00	0.33	0.00	0.17	0.50	1.00	16.67%
Total	0.17	0.42	0.67	0.92	0.25	0.67	2.92		
Precision	100.00%	0.00%	37.50%	18.18%	0.00%	25.00%			

A very small percentage of participants in the Edges group were able to identify room types correctly, as shown in Table 4.8.

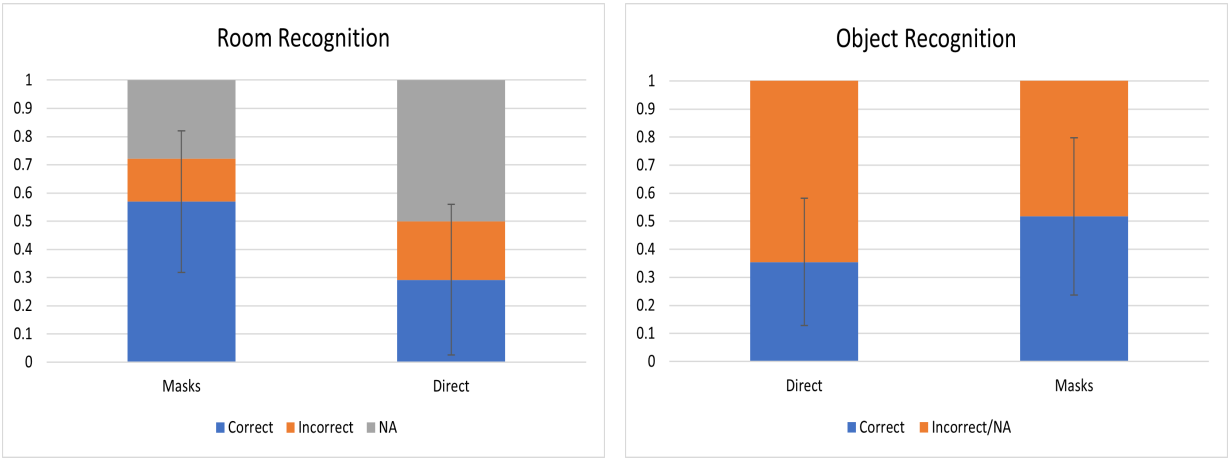
Overall, as derived from the above confusion matrices (Tables 4.6, 4.7, 4.8) the average value of precision for the enhancement group was 79% while the direct method was a mere 50% and the edges method was significantly lower. This is a sign that there were fewer false positives in our enhancement group and that there is a higher probability of correctly identifying the room type using the enhancement method than the direct or edge methods.

4.3 Experiment 2 (VR) results

A different set of images with different background colours was used to create a more challenging set of images and make the difference between the two groups clearer.

Table 4.9: Overall results for VR experiments

	Correct	Incorrect	NA	0	1	2	3	4	5
Enhancement	56.94%	15.28%	27.78%	28.21%	15.38%	19.23%	35.90%	51.28%	57.69%
Direct	29.17%	20.83%	50.00%	38.46%	8.97%	29.49%	38.46%	37.18%	55.13%



(a) (b)

Figure 4.6: Overall results for room identification and object recognition task using Virtual Reality. Percentage of correct, incorrect and not answered responses.

As shown in Figure 4.6, the percentage of correct responses for room types and objects within the room is higher in the enhancement group (Masks), with a p-value less than 0.01, meaning that the difference is statistically significant.

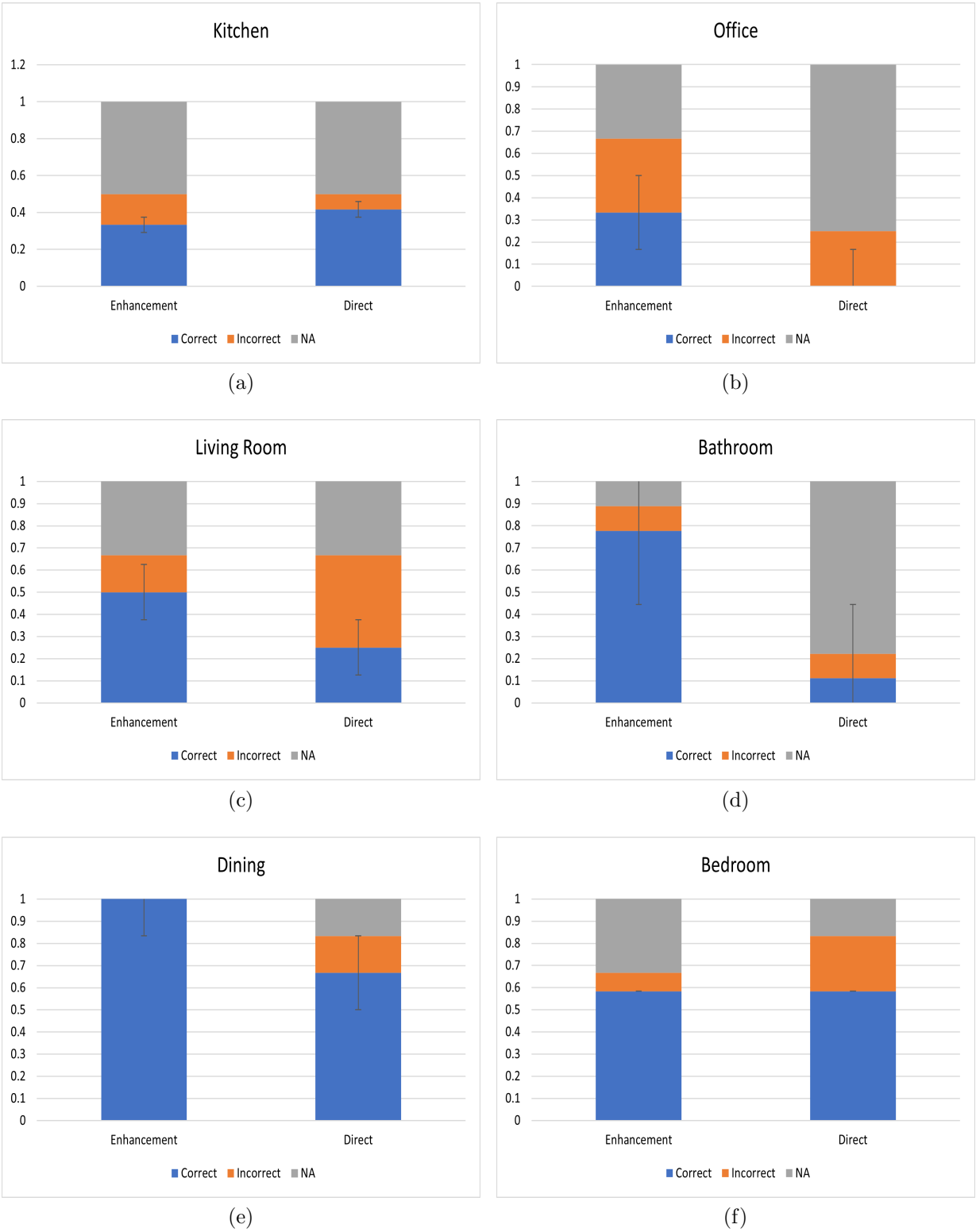


Figure 4.7: Room identification results for each room-type.

Room identification results for each room type. Higher scores in correct responses indicate that subjects were able to recognize the type of room in each test image. Higher ratios of non-responses indicate that subjects were not able to recognize the type of room in each image. According to the results displayed in Fig.4.7, the largest difference between the two groups is clearer in the bathroom and office room types. There are some room types where Direct performs better, but the difference is so small it could be due to chance.

Table 4.10: Confusion matrix results of room type recognition by using the Enhancement method.

Actual/Predicted	**Bathroom**	**Kitchen**	**Bedroom**	**Living Room**	**Office**	**Dining Room**	**NA**	**Total**	**Recall**
Bathroom	0.78	0.00	0.06	0.00	0.06	0.00	0.00	0.89	87.50%
Kitchen	0.08	0.42	0.08	0.00	0.00	0.00	0.50	1.08	38.46%
Bedroom	0.08	0.00	0.58	0.00	0.00	0.00	0.33	1.00	58.33%
Living Room	0.00	0.00	0.17	0.50	0.00	0.00	0.33	1.00	50.00%
Office	0.00	0.08	0.00	0.00	0.33	0.17	0.33	0.92	36.36%
Dining Room	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	100.00%
Total	0.94	0.50	0.89	0.50	0.39	1.17	1.50		
Precision	82.35%	83.33%	65.63%	100.00%	85.71%	85.71%			

The results in Table 4.10, which belongs to the Enhancement group, show that some participants confused room types, perhaps due to the similarity of the objects present in them. For example, a larger percentage of participants confused the office with the dining room because of the presence of tables and chairs. Some participants confused the living room with the bedroom, mistakenly identifying the couch as a bed. However, the diagonal (True Positives) usually had the highest percentage of responses, and the values of the diagonal in the enhancement table are usually higher than in Table 4.11, which belongs to the Direct group.

Table 4.11: Confusion matrix results of room type recognition by using the Direct method.

Actual/Predicted	**Bathroom**	**Kitchen**	**Bedroom**	**Living Room**	**Office**	**Dining Room**	**NA**	**Total**	**Recall**
Bathroom	0.22	0.11	0.00	0.00	0.00	0.00	0.25	0.58	38.10%
Kitchen	0.00	0.42	0.08	0.00	0.00	0.00	0.50	1.00	41.67%
Bedroom	0.00	0.08	0.50	0.00	0.17	0.00	0.08	0.83	60.00%
Living Room	0.25	0.00	0.08	0.17	0.08	0.00	0.33	0.92	18.18%
Office	0.00	0.00	0.00	0.00	0.00	0.17	0.75	0.92	0.00%
Dining Room	0.00	0.17	0.00	0.00	0.00	0.67	0.17	1.00	66.67%
Total	0.47	0.78	0.67	0.17	0.25	0.83	2.08		
Precision	47.06%	53.57%	75.00%	100.00%	0.00%	80.00%			

4.4 Experiment 3 (Real-time) results

The maximum allowed time for this experiment for each participant is 15 minutes. Results show that the enhancement group was generally faster at completing the experiment (identifying the objects and navigating through the room), as shown in Figure 4.8.

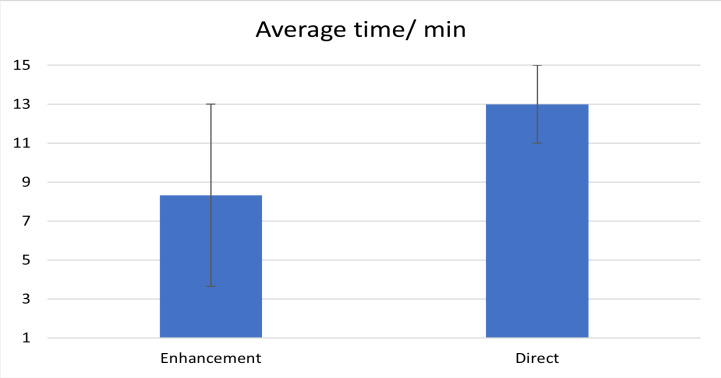


Figure 4.8: Average total time taken for each group.

Overall, all participants, regardless of the group, were able to identify the main features of the room, such as computer screens, keyboards, and mice. This could be slightly biased since all of them are university students, and they might have some expectations regarding the room on the campus. However, when it comes to the finer details, such as identifying a remote control, watch, potted plant, or water bottle, the enhancement group tends to do better, as shown in Figure 4.9.

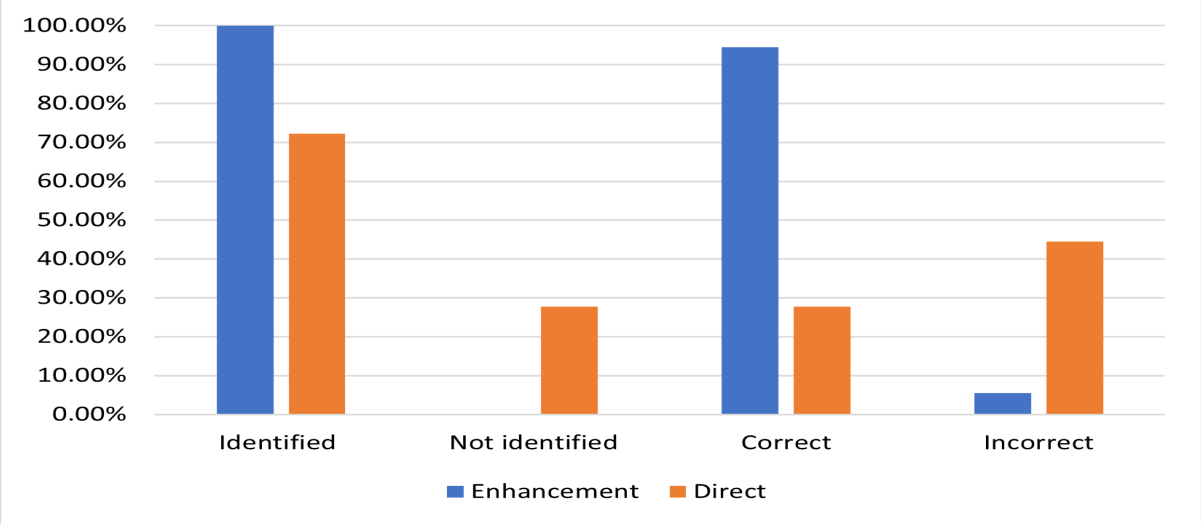


Figure 4.9: Average percentage of identified, unidentified, correctly and incorrectly identified objects.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The invention of visual prostheses has the potential to transform the lives of blind people. In this research, we mainly focused on improving indoor scene understanding for visually impaired patients. This was achieved by investigating the effects of different image processing techniques and using machine learning [65] to enhance the input to the visual prosthesis. We focused on six different indoor scenes: living room, bedroom, bathroom, kitchen, dining room, and office. To verify our method, we tested it against the traditional image processing techniques (edge detection method) and we also compared it to the direct method, where no image processing is used. Then, we tested the effect of our method by simulating prosthetic vision. This was done using a regular computer screen, a recorded virtual reality environment, and a real-time virtual reality environment, with each experiment having at least 12 participants of varying ages and genders. Our results demonstrate that our method is well suited for indoor scene understanding over traditional image processing methods used in visual prostheses. The key finding of our current results is that, with only a few significant objects in the scene, it is possible to obtain a good understanding of the type of room, even in the presence of clutter. This work can be used to help visually impaired people significantly improve their ability to adapt to the surrounding environment [9].

5.2 Future Work

Phosphene simulation is a challenging topic due to the limited capabilities of retinal implants. For this research, we have a possible phosphene simulation that would have been available if retinal implants were advanced enough, but in reality, the visual field is very narrow, and phosphenes are not perfectly located in the visual field corresponding to a grayscale pixel, and their size and form tend to differ [60]. Another limitation reported by some patients is phosphene dropout [75]. In the future, it might be worthwhile to try

different representations of phosphene simulations to investigate the extent of the effect they have on visual interpretation. Furthermore, indoor scenes can be very complex and there are a variety of different objects that can be present depending on the room type. It would be better to include more scenes and fine-tune machine learning models such that they would recognize all the objects that are likely to be present within these scenes. Due to the limited types of rooms available, the real-time experiment was only conducted in computer labs. However, undertaking the experiment in more room types would have added to the results and confirmed our findings. Overall, we can affirm that the comprehension of indoor scenes can be obtained with just a few sets of elements represented in the environment.

Appendix

Appendix A

Lists

List of Figures

List of Tables

Bibliography

[1] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Josechu Guerrero. Indoor scenes understanding for visual prosthesis with fully convolutional networks. pages 218–225, 01 2019.

[2] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Jose J Guerrero. Semantic and structural image segmentation for prosthetic vision. *Plos one*, 15(1):e0227677, 2020.

[3] World Health Organization. Blindness and vision impairment, 1890.

[4] Neil M Schultz, Shweta Bhardwaj, Claudia Barclay, Luis Gaspar, and Jason Schwartz. Global burden of dry age-related macular degeneration: a targeted literature review. *Clinical therapeutics*, 43(10):1792–1818, 2021.

[5] U Rajendra Acharya, Muthu Rama Krishnan Mookiah, Joel En Wei Koh, Jen Hong Tan, Kevin Noronha, Sulatha Bhandary, A. Rao, Yuki Hagiwara, Kuang Chua, and Augustinus Laude. Novel risk index for the identification of age-related macular degeneration using radon transform and dwt features. *Computers in Biology and Medicine*, 73:131–140, 06 2016.

[6] National Eye Institute. Age-related macular degeneration (amd), 2020.

[7] Cleveland Clinic. Age-related macular degeneration, 2020.

[8] American Academy of Opthamology. Have amd? save your sight with an amsler grid, 2020.

[9] Allen Pelletier, Jeremy Thomas, and Fawwaz Shaw. Vision loss in older persons. *American family physician*, 79:963–70, 07 2009.

[10] erkut küçük. Causes of blindness and moderate-severe visual impairment in niğde, central anatolia, turkey. *Erciyes Medical Journal*, 41, 11 2019.

[11] Xun Zhang, Ali Tohari, Fabio Marcheggiani, Xinzhi Zhou, James Reilly, Luca Tiano, and Xinhua Shu. Therapeutic potential of co-enzyme q10 in retinal diseases. *Current Medicinal Chemistry*, 24, 08 2017.

[12] Kristin Galetta, Peter Calabresi, Elliot Frohman, and Laura Balcer. Optical coherence tomography (oct): Imaging the visual pathway as a model for neurodegeneration. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, 8:117–32, 01 2011.

[13] Maximilian Treder, Jost Lauermann, and Nicole Eter. Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 256, 11 2018.

[14] Sharon Kingman. Glaucoma is second leading cause of blindness globally. *Bulletin of the World Health Organization*, 82:887–8, 12 2004.

[15] Qaisar Abbas. Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. *International Journal of Advanced Computer Science and Applications*, 8, 01 2017.

[16] National Eye Institute. Types of glaucoma, 2021.

[17] Ankur Sinha, Shalini Mohan, Viney Gupta, and Ramanjit Sihota. Gonioscopy. *DOS Times*, 10:322–328, 03 2005.

[18] Shalini Mohan, Anand Aggarwal, Tanuj Dada, Murugesan Vanathi, and Anita Panda. Pachymetry: A review. *DOS Times*, 12:19–28, 04 2007.

[19] David Huang, Eric Swanson, Charles Lin, Joel Schuman, William Stinson, Warren Chang, Michael Hee, Thomas Flotte, Kenton Gregory, Carmen Puliafito, and James Fujimoto. Optical coherence tomography. *Science*, 254:1178, 12 1991.

[20] Fabio Kanadani, TCA Moreira, L Campos, M Vianello, J Corradi, S Dorairaj, ALA Freitas, and Robert Ritch. A new provocative test for glaucoma. *Journal of Current Glaucoma Practice with DVD*, 10:1–3, 04 2016.

[21] Chris Johnson. *Visual Fields: Visual Field Test Strategies*, pages 145–151. 07 2016.

[22] Lance Doucette and Michael Walter. Prostaglandins in the eye: Function, expression, and roles in glaucoma. *Ophthalmic Genetics*, 38:1–9, 04 2016.

[23] V.P. Erichev, Sergey Petrov, Andrew Volzhanin, D.M. Safonova, T.V. Yaremenko, and S.A. Kazaryan. Alpha-adrenergic receptor agonists in terms of modern views on glaucoma monitoring and treatment. *Clinical ophthalmology*, 19:87–91, 07 2019.

[24] N. Moura-Coelho, Joana Ferreira, Carolina Bruxelas, Marco Dutra Medeiros, João Cunha, and Rita Pinto Proença. Rho kinase inhibitors—a review on the physiology and clinical use in ophthalmology. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 257:1–17, 06 2019.

[25] John Larkin. Laser treatment for glaucoma. *The Lancet*, 396:754, 09 2020.

[26] Mitsuru Nakazawa. Glaucoma filtering surgery. 65:1572–1575, 10 2011.

[27] Parul Singh, Krishna Kuldeep, Manoj Tyagi, ParmeshwariD Sharma, and Yogesh Kumar. Glaucoma drainage devices. *Journal of Clinical Ophthalmology and Research*, 1:77, 01 2013.

[28] Nick Astbury. Alternative eye care. *The British journal of ophthalmology*, 85:767–8, 08 2001.

[29] Lloyd Aiello, T Gardner, G King, G Blankenship, Jerry Cavallerano, Frederick Ferris, and R Klein. Diabetic retinopathy. *Diabetes care*, 21:143–56, 01 1998.

[30] Dr Sujit Murade. Diabetic-retinopathy, 1890.

[31] Jost B Jonas and Charumathi Sabanayagam. *Epidemiology and Risk Factors for Diabetic Retinopathy*, pages 20–37. 01 2019.

[32] Rwan Radi, Elaf Damanhouri, and Muhammad Siddiqui. Diabetic retinopathy causes, symptoms, complications: a review. *International Journal of Medicine in Developing Countries*, page 1, 01 2021.

[33] J Lock and Kenneth Fong. Retinal laser photocoagulation. *The Medical journal of Malaysia*, 65:88–94; quiz 95, 03 2010.

[34] Jennifer Evans and Gianni Virgili. Anti-vegf drugs: Evidence for effectiveness. *Community eye health / International Centre for Eye Health*, 27:48, 12 2014.

[35] Daniel Brănişteanu, A. Bilha, and Andreea Moraru. Vitrectomy surgery of diabetic retinopathy complications. *Romanian Journal of Ophthalmology*, 60:31–36, 01 2016.

[36] Mohammad Hossein Maghami, Amir Sodagar, Alireza Zabihian, and Farzad Asgarian. Implantable biomedical devices. 09 2012.

[37] Peter Walter. *Retinal Implants*, pages 1–11. 06 2006.

[38] Dr Sujit Murade. artificial$_s ight$, 1890.

[39] Jia-Wei Yang, Zih-Yu Yu, Sheng-Jen Cheng, Johnson Chung, Xiao Liu, Chung-Yu Wu, Shien-Fong Lin, and Guan-Yu Chen. Graphene oxide–based nanomaterials: An insight into retinal prosthesis. *International Journal of Molecular Sciences*, 21:2957, 04 2020.

[40] Shuai Niu, Jinhai Niu, Yifei Liu, Yang Zhou, and Qiushi Ren. Implantable system for optic nerve visual prosthesis. pages 1503 – 1506, 06 2007.

[41] Ashish Tiwari and R.H. Talwekar. Review on progressive development of cmos imagers for visual prosthesis and new aspects. *Journal of Advanced Research in Dynamical and Control Systems*, 9:1334–1348, 01 2017.

[42] Soroush Niketeghad and Nader Pouratian. Brain machine interfaces for vision restoration: The current state of cortical visual prosthetics. *Neurotherapeutics*, 16, 09 2018.

[43] Margee J Kyada, Nathaniel J Killian, and John S Pezaris. Thalamic visual prosthesis project. In *Artificial Vision*, pages 177–189. Springer, 2017.

[44] Hieu T. Nguyen, Siva M. Tangutooru, Corey M. Rountree, Andrew J. Kantzos, Faris Tarlochan, W. Jong Yoon, and John B. Troy. Thalamic visual prosthesis. *IEEE Transactions on Biomedical Engineering*, 63(8):1573–1580, 2016.

[45] Istvan Bokkon. Phosphene phenomenon: A new concept. *Bio Systems*, 92:168–74, 06 2008.

[46] Spencer Chen, Gregg Suaning, John Morley, and Nigel Lovell. Simulating prosthetic vision: I. visual models of phosphenes. *Vision research*, 49:1493–506, 07 2009.

[47] Giles S Brindley and Walpole S Lewin. The sensations produced by electrical stimulation of the visual cortex. *The Journal of physiology*, 196(2):479–493, 1968.

[48] Giles S Brindley and David N Rushton. Symposium-prosthetic aids for blind-implanted stimulators of visual cortex as visual prosthetic devices. *TRANSACTIONS AMERICAN ACADEMY OF OPHTHALMOLOGY AND OTOLARYNGOLOGY*, 78(5):O741–O745, 1974.

[49] William H Dobelle, Michael G Mladejovsky, and JP Girvin. Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science*, 183(4123):440–444, 1974.

[50] BS Everitt and DN Rushton. A method for plotting the optimum positions of an array of cortical electrical phosphenes. *Biometrics*, pages 399–410, 1978.

[51] DN Rushton and GS Brindley. Properties of cortical electrical phosphenes. pages 574–593, 1978.

[52] Jean Delbeke, Medhy Oozeer, and Claude Veraart. Position, size and luminosity of phosphenes generated by direct optic nerve stimulation. *Vision research*, 43(9):1091–1102, 2003.

[53] Spencer C Chen, Gregg J Suaning, John W Morley, and Nigel H Lovell. Simulating prosthetic vision: I. visual models of phosphenes. *Vision research*, 49(12):1493–1506, 2009.

[54] Må E Brelén, F Duret, Benoît Gérard, Jean Delbeke, and Claude Veraart. Creating a meaningful visual perception in blind volunteers by optic nerve stimulation. *Journal of neural engineering*, 2(1):S22, 2005.

[55] Manjunatha Mahadevappa, James D Weiland, Douglas Yanai, Ione Fine, Robert J Greenberg, and Mark S Humayun. Perceptual thresholds and electrode impedance in three retinal prosthesis subjects. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):201–206, 2005.

[56] G Richard, R Hornig, M Keserü, and M Feucht. Chronic epiretinal chip implant in blind patients with retinitis pigmentosa: long-term clinical results. *Investigative Ophthalmology & Visual Science*, 48(13):666–666, 2007.

[57] E Zrenner, D Besch, KU Bartz-Schmidt, F Gekeler, VP Gabel, C Kuttenkeuler, H Sachs, H Sailer, B Wilhelm, and R Wilke. Subretinal chronic multi–electrode arrays implanted in blind patients. *Investigative Ophthalmology & Visual Science*, 47(13):1538–1538, 2006.

[58] E Zrenner, R Wilke, T Zabel, H Sachs, K Bartz-Schmidt, F Gekeler, B Wilhelm, U Greppmaier, A Stett, SUBRET Study Group, et al. Psychometric analysis of visual sensations mediated by subretinal microelectrode arrays implanted into blind retinitis pigmentosa patients. *Investigative Ophthalmology & Visual Science*, 48(13):659–659, 2007.

[59] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Josechu Guerrero. Structural and object detection for phosphene images, 09 2018.

[60] Michael Beyeler, Devyani Nanduri, James Weiland, Ariel Rokem, Geoffrey Boynton, and Ione Fine. A model of ganglion axon pathways accounts for percepts elicited by retinal implants. 10 2018.

[61] Jeong-ah Kim, Ju-Yeong Sung, and Se-ho Park. Comparison of faster-rcnn, yolo, and ssd for real-time vehicle type recognition. In *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pages 1–4, 2020.

[62] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[63] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. pages 3039–3048, 07 2017.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 06 2015.

[65] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. pages 2980–2988, 10 2017.

[66] Shubham Shinde, Ashwin Kothari, and Vikram Gupta. Yolo based human action recognition and localization. *Procedia Computer Science*, 133:831–838, 01 2018.

[67] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[68] Tech Comp. artificial$_s ight$, 1890.

[69] Wen Dai, Jiaming Na, Nan Huang, Guanghui Hu, Xin Yang, Guoan Tang, LiYang Xiong, and Fayuan Li. Integrated edge detection and terrain analysis for agricultural terrace delineation from remote sensing images. *International Journal of Geographical Information Science*, 34:1–20, 08 2019.

[70] Satbir Kaur and Ishpreet Singh Virk. Comparison between edge detection techniques. *International Journal of Computer Applications*, 145:15–18, 07 2016.

[71] Mohammad Awrangjeb and Guojun Lu. An improved curvature scale-space corner detector and a robust corner matching approach for transformed image identification. *Image Processing, IEEE Transactions on*, 17:2425 – 2441, 01 2009.

[72] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Josechu Guerrero. Structural and object detection for phosphene images, 09 2018.

[73] Mouna Afif, Riadh Ayachi, Yahia Said, and Mohamed Atri. Deep learning based application for indoor scene recognition. *Neural Processing Letters*, 51, 06 2020.

[74] Michael Beyeler, Geoffrey M Boynton, Ione Fine, and Ariel Rokem. pulse2percept: A python-based simulation framework for bionic vision. *BioRxiv*, page 148015, 2017.

[75] Yanyu Lu, Jing Wang, Hao Wu, Liming Li, Xun Cao, and Chai Xinyu. Recognition of objects in simulated irregular phosphene maps for an epiretinal prosthesis. *Artificial organs*, 38, 10 2013.