

Image Perception Enhancement in Prosthetic Vision

vorgelegt von
M.Sc.-Ing.
Reham Hossam Elnabawy

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
-Dr.-Ing.-
genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Marc Alexa

Gutachter: Prof. Dr.-Ing. Olaf Hellwich

Gutachter: Assoc. Prof. Seif Eldawlatly

Gutachter: Assistant Prof. John Pezaris

Gutachter: Prof. Walid Al-Atabany

Tag der wissenschaftlichen Aussprache: 12. Januar 2023

Berlin 2023

Zusammenfassung

Das Sehen ist der wichtigste Sinn, ohne den ein Mensch nicht unabhängig leben kann. Sehprothesen wurden als viel versprechende Lösung vorgeschlagen, um Menschen, die ihre Sehkraft verloren haben, zu helfen und ihnen ein teilweises Sehvermögen zu geben. Sie nutzen die funktionellen Teile des Auges und des Gehirns, um eine teilweise Wiederherstellung des Sehvermögens zu erzielen. Trotz anfänglicher Erfolge gibt es einige Probleme hinsichtlich der Fähigkeit der mit Sehprothesen ausgestatteten Patienten, Objekte richtig zu erkennen und zu lokalisieren. In dieser Arbeit stellen wir eine Reihe von Lösungen vor, um die Fähigkeiten von Sehprothesenträgern zu verbessern, Objekte richtig zu erkennen und zu lokalisieren. Techniken der Bildverarbeitung und des Deep Learning zur vereinfachten Nutzung von Sehprothesen werden entwickelt. Eine entscheidende Herausforderung, die bei einer typischen Sehprothese auftritt, ist der Ausfall von einzelnen Elektroden im Laufe der Zeit der Benutzung. Zur Bewältigung dieser Herausforderung wird eine optimale Lösung vorgeschlagen, indem das Objekt des Interesses an eine Position des Implantats im Gesichtsfeld des Benutzers verschoben wird, so dass das die Fläche des Objektsegments von möglichst wenigen ausfallenden Elektroden betroffen ist. Eine weitere Herausforderung für die Nutzer von Sehprothesen ist die Komplexität der betrachteten Szene, die aufgrund der geringen räumlichen und radiometrischen Auflösung, die mit Sehprothesen möglich ist, kaum erkennbar ist. Dementsprechend schlagen wir die Verwendung von clip art-Darstellungen der abgebildeten Objekte anstelle der realen Fotos vor, um die Erkennung beliebiger Objekte zu erleichtern. Wir schlagen die Verwendung von You Only Look Once (YOLO) vor, einem Modell für tiefes Lernen, um die clip art Darstellung abzurufen, die einem Objekt entspricht, das, auch mittels YOLO, in einem hochauflösenden Foto erkannt wurde. Darüber hinaus wird ein auf Deep Learning basierender Ansatz mit generativen adversarialen Netzwerken (GANs), genannt PVGAN, vorgeschlagen, um clip art-Bilder aus gegebenen hochauflösenden Fotos zu generieren, um eine bessere und einfachere Wahrnehmung der Bilder in einer Sehprothese zu ermöglichen. Schließlich kombinieren wir drei Methoden zur Verbesserung der Objekterkennung und lokalisierung mit den GAN-generierten Cliparts, nämlich Kantenschärfung, Eckenschärfung und Dropout. Der kombinierte Ansatz wird in einer Gemischte-Realität-Umgebung getestet, um die visuelle Darstellung zu simulieren, die von Prothesenbenutzern wahrgenommen wird. Eine Reihe von Experimenten zur Simulation des Sehvermögens von Prothesenträgern wurde an normalsichtigen Teilnehmern durchgeführt, um die Wirksamkeit der vorgeschlagenen Ansätze sowohl auf dem Computerbildschirm als auch in der gemischte Realität zu ermitteln. Die Ergebnisse zeigen, dass die Anwendung der vorgeschlagenen Techniken die Fähigkeit von Menschen verbessert, Objekte korrekt zu erkennen und zu lokalisieren. Dies könnte Nutzern von Sehprothesen ermöglichen, ihr Selbstvertrauen und ihre Unabhängigkeit wiederzuerlangen.

Abstract

Vision is the most crucial sense that a human being cannot live independently without. Visual prosthesis has been proposed as a promising solution to restore partial vision to people who lost their vision. Visual prostheses exploit the functional parts in the eye and the brain to enable partial restoration of vision. Despite its initial success, some challenges arise that hinder the ability of the patients implanted with visual prostheses to correctly recognize and localize objects. In this thesis, we introduce a variety of solutions to enhance the ability of visual prostheses users to correctly recognize and localize objects. Image processing and deep learning techniques are proposed in this thesis to simplify and better represent objects for visual prostheses users. One crucial challenge that arises in a typical visual prosthetic device is electrodes dropout, where some electrodes malfunction throughout time. To address this challenge, an optimal solution is proposed by translating the object of interest to a location in the visual field of the user such that the minimum amount of dropout exists. Another challenge that visual prostheses users face is the complexity of the viewed scene that is barely recognizable due to the low spatial and radiometric resolutions available through visual prostheses. Accordingly, we propose the utilization of clip art representation of images instead of the actual real photo to ease the recognition of any arbitrary object. We propose the use of You Only Look Once (YOLO), a deep learning model to retrieve the clip art that corresponds to an object detected by YOLO in a high-resolution photo. In addition, a deep learning-based approach using Generative Adversarial Networks (GANs), named PVGAN, is proposed to generate clip art images from given high-resolution photos to allow better and easier perception of the images in a visual prosthetic device. Finally, we combine three enhancement techniques with the GAN-generated clip art which are edge sharpening, corners sharpening and dropout handling to enhance object recognition and localization. The combined approach is tested in a mixed reality environment to simulate the visual representation perceived by visual prostheses users. A number of prosthetic vision simulation experiments were conducted on normally/correctly sighted participants to measure the efficacy of the proposed approaches using both computer screen and mixed reality. The results demonstrate that the usage of the proposed techniques enhances the ability of people to correctly recognize and localize objects. This could allow visual prostheses users to regain back their confidence and independence.

Dedicated to ...

Acknowledgements

I would like to express my sincere gratitude to my supervisors for the continuous support throughout my doctoral thesis. I would like to express my gratitude to Prof. Olaf Hellwich for his constructive feedback and for the exciting environment that shaped every meeting. Moreover, I would like to extend my gratitude to Assoc. Prof. Seif Eldawlatly for his continuous guidance and continuously giving constructive feedback, while paying attention to the smallest details to ensure that everything is on the right track. Furthermore, I would like to thank Prof. Slim Abdennadher for his recommendations in regards to the flow of the doctoral thesis. I would like to thank Yasmin Abdelghaffar for the artwork shown in a figure utilized in this thesis. In addition, I would like to thank all the volunteers who participated in the experiments presented in this thesis. Finally, this thesis would not have been finished without the support of my mother and my brother whom without them, this thesis would never be completed in the best possible way.

Table of Contents

Title Page	i
Zusammenfassung	iii
Abstract	v
List of Figures	xv
List of Tables	xix
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	2
1.3 Thesis Contributions	2
1.3.1 Dropout Handling	2
1.3.2 Scene Simplification	3
1.3.2.1 YOLO-based Object Simplification Approach for Visual Pros- theses	3
1.3.2.2 A Generative Adversarial Network for Object Simplification in Prosthetic Vision	3
1.3.3 Real-Time Mixed Reality Simulation	4
1.4 Thesis Outline	4
2 Background	5
2.1 Overview of the Eye and the Brain	5
2.2 The Retina	6
2.3 Visual Prosthesis	7
2.3.1 Nature of Stimulation Signals	9
2.3.2 Phosphenes	10
2.3.2.1 Phosphene Simulation	10
2.4 Visual Prostheses Types	12
2.4.1 Electrode-Based Visual Prostheses	12
2.4.1.1 Comparison of Different Approaches	13
2.4.2 Optogenetic-Based Visual Prostheses	13
2.5 Visual Prostheses User Experience and Challenges	14
2.6 Image Processing Techniques	16

TABLE OF CONTENTS

2.6.1	Techniques Used	16
2.6.1.1	Histogram of Oriented Gradients	16
2.6.1.2	Otsu Thresholding	16
2.6.1.3	Dilation	18
2.6.1.4	Skeletonization	18
2.6.1.5	Connected Components Labelling	19
2.6.1.6	Features from Accelerated Segment Test	21
2.6.1.7	Contrast Limited Adaptive Histogram Equalization	22
2.6.1.8	Median Filter	23
2.6.1.9	Wiener Filter	23
2.6.2	Related Work	24
2.7	Deep Learning Techniques	25
2.7.1	Techniques Used	25
2.7.1.1	Convolutional Neural Network	25
2.7.1.2	You Only Look Once	28
2.7.1.3	Generative Adversarial Network	31
2.7.2	Related Work	37
3	Image Enhancement and Phosphene Simulation	39
3.1	Image Enhancement	39
3.2	Phosphene Simulation	42
4	Electrode Dropout Handling	47
4.1	Introduction	47
4.2	Methods	47
4.2.1	Image Pre-processing	47
4.2.2	Phosphene Simulation	48
4.2.3	Dropout Handling Approach	48
4.2.4	Experimental Design and Procedure	48
4.2.5	Evaluation Metrics	49
4.3	Results	49
4.3.1	Dropout Simulation and Handling	49
4.3.2	Performance Evaluation	50
4.4	Conclusion	54
5	Scene Simplification	55
5.1	YOLO-Based Scene Simplification	55
5.1.1	Introduction	55
5.1.2	Methods	56
5.1.2.1	Experimental Design and Procedure	56
5.1.2.2	YOLO-based Clip Art Retrieval	57
5.1.2.3	Image Pre-processing	57
5.1.2.4	Phosphene Simulation	57
5.1.2.5	Evaluation Metrics	58

5.1.3	Results	58
5.1.3.1	Object Recognition Enhancement	58
5.1.3.2	Performance Evaluation	59
5.1.4	Conclusion	61
5.2	A GAN Application for Clip Art Generation	61
5.2.1	Introduction	61
5.2.2	PVGAN Model Overview	62
5.2.3	Training Datasets	64
5.2.4	Pre-processing of Training Input Data	66
5.2.5	Phosphene Simulation	67
5.2.5.1	The Scoreboard Model	67
5.2.5.2	The Axon Map Model	67
5.2.6	ClipArtGAN Evaluation Experiments	67
5.2.7	PVGAN Experiments Overview and Setup	68
5.2.7.1	Scoreboard Phosphene Simulation Experiments	69
5.2.7.2	Axon Map Phosphene Simulation Experiments	69
5.2.8	PVGAN Experiments Evaluation Metrics	70
5.3	Results	70
5.3.1	Training Datasets Pre-Processing Overview	70
5.3.2	Training Datasets Qualitative Evaluation	72
5.3.3	ClipArtGAN Experimental Evaluation	75
5.3.4	Phosphenes Simulation Outcome	77
5.3.5	PVGAN Experimental Results	81
5.3.5.1	Scoreboard Model Results	81
5.3.5.2	Axon Map Model Results	83
5.3.6	Performance Analysis	86
5.3.6.1	ClipArtGAN	86
5.3.6.2	PVGAN	87
5.3.7	Conclusion	88
6	Mixed Reality Real-Time Simulation	90
6.1	Introduction	90
6.2	Methods	90
6.2.1	System Overview	90
6.2.2	Phosphene Simulation	92
6.2.3	Proposed Enhancement Techniques	93
6.2.3.1	Clip Art Representation	93
6.2.3.2	Edge Enhancement	93
6.2.3.3	Corner Enhancement	93
6.2.4	Tools Used	94
6.2.5	Experimental Setup	94
6.2.5.1	Single Object Recognition and Localization Experiments	96
6.2.5.2	Multiple Objects Recognition and Localization Experiments	97

TABLE OF CONTENTS

6.2.5.3	Object Recognition and Localization during Navigation Experiments	98
6.2.6	Evaluation Metrics	98
6.3	Results	99
6.3.1	Enhancement Techniques and Phosphenes Simulation Outcome	99
6.3.2	Experimental Results	101
6.3.2.1	Single Object Experiments	101
6.3.2.2	Multiple Objects' Experiments	107
6.3.2.3	Navigation Experiments	110
6.4	Analysis of the Utilization of the Enhancement Techniques	113
6.5	Conclusion	115
7	Conclusion and Future Work	116
7.1	Conclusion	116
7.2	Future Work	117
	References	119

List of Figures

2.1	The visual projection pathway (Adapted from [8]).	6
2.2	Human eye retina anatomy (Adapted from [9]).	7
2.3	Visual prosthesis. (a) Implanted components of the system. (b) External Components of the system. (c) Electrode array implanted in the retina. (Adapted from [19, 20]).	9
2.4	Phosphenes sample. (a) Input image. (b) Phosphene simulation. (Adapted from [25]).	10
2.5	Squared and hexagonal grids. (a) Original image. (b) Squared-grid phosphene simulation. (c) Hexagonal-grid phosphene simulation.	11
2.6	Binary and grayscale scoreboard phosphene simulation. (a) Original image. (b) Binary scoreboard phosphene simulation. (c) Grayscale scoreboard phosphene simulation.	12
2.7	Dropout rates of (a) 10%, (b) 20% and (c) 30%, respectively.	12
2.8	Face sample in simulated prosthetic vision with no dropout and with dropout added (Adapted from [44]).	15
2.9	Dilation applied on input image. (a) Input image. (b) Cross structuring element. (c) Dilated image. (Adapted from [48]).	18
2.10	Skeletonization sample. (a) Original image. (b) Skeletonized image. (Adapted from [50]).	19
2.11	Connected components labelling. (a) 4-connectivity. (b) S8-connectivity. (Adapted from [51]).	20
2.12	Connected components labelling algorithm. (a) Binary image. (b) Three objects are detected by 4-connectivity component labelling. (c) Two objects are detected by 8-connectivity component labelling. (Adapted from [52]).	20
2.13	The 16 values surrounding pixel p stored in vector form (Adapted from [54]).	21
2.14	An overview of the training procedure and a convolutional neural network (CNN) architecture (Adapted from [62]).	26
2.15	Common activation methods for neural networks. (a) rectified linear unit (ReLU). (b) sigmoid. (c) hyperbolic tangent (tanh). (Adapted from [62]).	26
2.16	Underfitting versus overfitting (Adapted from [62]).	27
2.17	YOLOv3 network architecture (Adapted from [68]).	29
2.18	GAN architecture. (a) Discriminator perspective. (b) Generator perspective. (Adapted from [70]).	32
2.19	CycleGAN Architecture (Adapted from [72]).	33

2.20	Pix2Pix GAN architecture (Adapted from [75]).	34
2.21	Generator and discriminator architectures. (a) U-Net Architecture used in the generator (Adapted from [80]). (b) PatchGAN classifier used in the discriminator (Adapted from [81]).	36
3.1	Experimental setup.	46
3.2	An overview of a real visual prosthesis system utilizing the proposed techniques.	46
4.1	Dropout simulation and handling for different dropout rates of 10%, 20% and 30%.	50
4.2	Results for 10% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level ($mean \pm std$). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	51
4.3	Results for 20% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level ($mean \pm std$). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	53
4.4	Results for 30% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level ($mean \pm std$). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	54
5.1	Block diagram of the proposed approach.	56
5.2	Phosphene simulation of each object for actual and clip art representations. The first group of subjects was presented with the phosphene simulation of the input photo, the focused objects and their zoomed in version. The second group of subjects was only presented with the phosphene simulation of the clip art representation only. The third group of subjects was presented with the phosphene simulation of the input photo, the focused objects and their clip art representation.	59
5.3	Performance of different groups of subjects in the task of recognizing the objects measured by (a) time to decision, (b) recognition accuracy, and (c) confidence level ($mean \pm std$). Blue bars represent using clip art for object simplification, while the red bar represent zooming on the actual object of interest. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	60
5.4	Block Diagram showing the flow of the proposed model. (a) ClipArt-GAN/PVGAN Training. (b) Testing the Model and Phosphene Simulation.	64
5.5	Clip Art pre-processing.	71
5.6	HOG best matching clip art selection.	72
5.7	Generated Clip Art versus Google Images Clip Art.	75
5.8	Scoreboard versus axon map phosphene simulation.	78
5.9	Results from scoreboard phosphene simulation for new test images from training classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level ($mean \pm std$). $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	82

5.10	Results from scoreboard phosphene simulation for test images from new classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (<i>mean</i> \pm <i>std</i>). $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	83
5.11	Results from axon map phosphene simulation for test images from training classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (<i>mean</i> \pm <i>std</i>). $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	85
5.12	Results from axon map phosphene simulation for test images from new classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (<i>mean</i> \pm <i>std</i>). $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.	86
6.1	Displayed Phosphene Simulation. (a) Sample Phosphene Simulation Perception in Experiments (1) Before Image Enhancement. (2) After Image Enhancement. (b) Visual Field Adjustment in Phosphene Simulation.	92
6.2	Experiments Set-Up. (a) Experimental Paradigm Timeline. (b) Single object Experiment Setup. (c) Multiple objects Experiment Setup. (d) Navigation Experiment Setup.	96
6.3	Single Objects Experiments. (a) Enhancement Techniques. (b) Phosphene Simulation of the 8 experiments.	100
6.4	Multiple Objects Experiments. (a) Control Group. (b) All Enhancement Techniques.	101
6.5	Navigation Experiments. (a) Control Group. (b) All Enhancement Techniques.	101
6.6	Single Objects Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. $*P < 0.05$, $**P < 1e - 04$, two-sample t-test.	103
6.7	Single Objects Recognition Performance for Each Displayed Object. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level.	104
6.8	Single Objects Experimental Results Statistics. (a) Grasping Attempt Time. (b) Grasping Accuracy. $*P < 0.05$, $**P < 1e - 04$, two-sample t-test.	106
6.9	Single Objects Grasping Performance for Each Displayed Object. (a) Grasping Attempt Time. (b) Grasping Accuracy.	107
6.10	Multiple Objects Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. $**P < 1e - 04$, two-sample t-test.	109
6.11	Multiple Objects Results for Each Object. (a) Recognition Time. (b) Recognition Accuracy (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. Blue Rectangles in All Figures Represent the Pairs of Objects that were Presented Together.	110
6.12	Navigation Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. $*P < 0.05$, $**P < 1e - 04$, two-sample t-test.	112
6.13	Navigation Results for Each Object. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy.	113

List of Tables

2.1	Comparison between different visual prosthesis approaches (Adapted from [37]).	13
3.1	Proposed image processing techniques with and without dilation.	41
3.2	Ideal phosphene simulation with no dilation.	42
5.1	Testing the CycleGAN model.	73
5.2	Testing the Pix2Pix model.	74
5.3	CycleGAN vs Pix2Pix GAN for ClipArt recognition Experiment.	76
5.4	Pix2Pix GAN vs Google Images for ClipArt generation Experiment.	77
5.5	Google Images recognizing generated Clip Art Experiment.	77
5.6	Scoreboard phosphene simulation.	79
5.7	Axon map phosphene simulation.	80

1

Introduction

1.1 Motivation

Blindness is a threat that impacts the lives of everyone who lost his/her vision. Some causes of blindness could be due to an accident or a disease such as cataracts which can be noticed when the eye lens gets less transparent, causing eyesight to appear cloudy. Moreover, age-related macular degeneration (AMD) causes blindness resulting in the loss of central vision. In addition, retinitis pigmentosa (RP) causes blindness as a result of losing the peripheral vision. The World Health Organization (WHO) estimates that more than 2.2 billion people worldwide suffer from blindness and visual impairment [1]. There are 39 million people around the globe, suffering from blindness [2]. This loss of vision leads to not only the loss of independence, but also the loss of the confidence for any blind person.

Multiple aids have been proposed to help blind people such as cane sticks, which were developed to help them to cross roads, navigate and avoid any possible obstacle. However, many accidents were reported despite the use of cane sticks, which demonstrates some disadvantages of using cane sticks [3]. Moreover, smart sticks were used to allow the blind patients to sense any obstacle in the range of the stick through sending vibrations as an alarm to the blind patient that there is an obstacle nearby [4]. Finally, applications of robotics were proposed to help in the safe mobility for blind patients through the reduction of cognitive load to provide them freedom [5].

Despite all the aforementioned aids, such aids only help blind patients, but do not restore their vision. Visual prostheses have been developed as devices that enable partial vision restoration to the blind. These devices interface with the visual system and send electrical pulses to activate visual system sites, thereby, inducing artificial visual percepts. The visual systems works by, first, transmitting incident light through the eye lens, then neural signals are converted by the retinal circuitry to electrical signals that are then, transmitted through the optic nerve to the thalamus and then the visual cortex where an image is rendered. Thus, losing the functionality of any of the aforementioned stages throughout this process, causes loss of vision (i.e., blindness). So, visual prosthesis has been a promising solution to give partial

restoration of vision to those people who lost their vision by targeting the functional sites within the visual system. A typical visual prosthesis device comprises a tiny video camera on the bridge of eyeglasses which captures images, and sends them to a video processing unit (VPU). This unit transforms the images into electrical signals and sends them back to an antenna on the glasses. The signals are then transferred wirelessly to the implanted electrodes.

Despite the potential benefit of visual prostheses, one drawback of the image perceived via visual prostheses is that the image is represented in relatively low-spatial and radiometric resolutions that affect the ability of the visual prosthetic user to recognize the objects in a scene. So accordingly, the motivation behind this thesis came from the fact that all people who were born with normal vision and then, lost their vision due to an infection, a disease, an accident or a genetic disorder, have their visual system fully developed. However, the loss of vision is due to the inability of their visual system to generate the needed signals. Visual prostheses compensate for these signals by artificially stimulating the visual system. Current prosthetic vision technology results in poor perceived image quality. Accordingly, it was thought to enhance the perceived image that is shown via visual prosthesis to enable blind people to be able to induce better perception of images to facilitate their daily life.

1.2 Problem Statement

Due to the lack of sufficient details in the image perceived using a typical visual prosthetic device, the independence level for visual prostheses users is relatively low affecting their daily activities. Therefore, it was thought to focus on the details of the perceived image and the main key points that could be easily determined by the users of visual prostheses devices. The research question is “How to enhance the images perception for visual prostheses users to regain their autonomy in performing daily activities?”

1.3 Thesis Contributions

This thesis proposes image processing and deep learning-based techniques to enhance image perception in visual prostheses.

1.3.1 Dropout Handling

Visual prostheses provide promising solution to the blind through partial restoration of their vision via electrical stimulation of the visual system. However, there are some challenges that hinder the ability of subjects implanted with visual prostheses to correctly identify an object. One of these challenges is electrode dropout; the malfunction of some electrodes resulting in consistently dark spots. In this thesis, we propose a dropout handling algorithm for better and faster identification of objects. In this algorithm, spots representing the object are translated to another location within the same image that has the minimum number of dropouts. Using simulated prosthetic vision, experiments were conducted to test the efficacy of our proposed algorithm. Electrode dropout rates of 10%, 20% and 30% were examined. Our results demonstrate significant increase in the object recognition accuracy of the participants,

reduction in the recognition time and increase in the recognition confidence level using the proposed approach compared to presenting the images without dropout handling.

1.3.2 Scene Simplification

1.3.2.1 YOLO-based Object Simplification Approach for Visual Prostheses

A picture is worth a thousand words. This is due to the significance of visualizing concepts rather than just representing them in textual form. While image representation is typically informative, simplifying image representations might be more informative when using images in visual prostheses. One simplified image representation of any object is to represent it in the form of a clip art. However, generating a clip art that corresponds to any arbitrary image is challenging. A challenge reported by visual prosthetic users is the difficulty of object recognition due to the low resolution of the images perceived through these devices. In this thesis, a deep learning-based approach combined with image pre-processing is proposed to allow visual prostheses' users to recognize objects in a certain scene. The approach simplifies the objects in the scene by displaying the objects in clip art form to enhance object recognition. These clip art images are retrieved by, first, identifying the objects in the scene using the You Only Look Once (YOLO) deep neural network. The clip art corresponding to each identified object is then retrieved via Google Images. The retrieved clip art is then displayed in simulation environment to mimic the perception of a typical visual prosthetic user. Three experiments were conducted to measure the success of the proposed approach using simulated prosthetic vision. Our results reveal a remarkable decrease in the recognition time, increase in the recognition accuracy and confidence level when using the clip art representation as opposed to using the actual images of the objects. These results demonstrate the utility of object simplification in enhancing the perception of images in prosthetic vision.

1.3.2.2 A Generative Adversarial Network for Object Simplification in Prosthetic Vision

We propose an approach to generate clip art images that correspond to any input image using Generative Adversarial Networks (GANs). The proposed approach is deep learning-based given the success deep learning demonstrated in image-to-image translation. The proposed deep learning model, PVGAN, utilizes GAN to convert a photo that contains only one object of interest to its corresponding clip art representation. We examined the efficacy of PVGAN prior to being used in visual prostheses. For simplicity, we named PVGAN as ClipArtGAN when it is used in any field other than visual prostheses field, where the nature and the number of training classes in PVGAN were different than that of ClipArtGAN to accommodate the objects that a visual prosthesis user might deal with in his/her daily life. Three evaluation experiments were conducted to examine ClipArtGAN before being used in visual prostheses. The first two experiments involved 12 participants to evaluate the outcomes of the GANs in comparison to the input photos, while the third experiment was performed using Google Images to measure the ability of clip art recognition. The results demonstrate that clip art images generated by ClipArtGAN are an accurate representation for the input photos in a simpler form of an image that can be used next in visual prostheses.

Accordingly, we utilized PVGAN, to enhance object recognition for the implanted patients by representing objects in the field of view based on a corresponding simplified clip art version. To assess the performance, two sets of simulated prosthetic vision experiments involving normally-sighted participants were performed. The two sets of experiments comprise different groups of approaches. The first and second groups comprised presenting simulation of the environment that a typical visual prosthetic user lives in, displaying the real images containing the actual high-resolution object, and presenting the real image followed by the clip art image, respectively. The other two groups were performed to evaluate the performance in the case of electrode dropout, where the third group comprised presenting only clip art images without electrode dropout, while the fourth group involved clip art images with electrode dropout in the simulation environment. Results demonstrate that representing the objects using clip art images generated by the PVGAN model results in a significant enhancement in the speed and confidence of the subjects in recognizing the objects. These results demonstrate the utility of using deep learning techniques, and GANs in particular, in enhancing the quality of images perceived using prosthetic vision.

1.3.3 Real-Time Mixed Reality Simulation

Edge sharpening, corners sharpening, and dropout handling are three enhancement techniques that we integrate with the GAN-generated clip art to improve object localization and recognition. The combined strategy is evaluated in a mixed reality environment to mimic the visual representation experienced by implanted visual prostheses users. Twelve experiments were conducted to measure the performance of the participants in object recognition and localization. The experiments involved single objects, multiple objects and navigation. The results demonstrate that the usage of dropout handling, clip art representation, edge enhancement and corners enhancement, gives higher accuracy, confidence level and less time for recognizing and grasping an object.

1.4 Thesis Outline

In this thesis, Chapter 2 discusses the background of human vision, visual prostheses, image processing techniques, computer vision techniques, deep learning techniques and literature review. Chapter 3 discusses the enhancement techniques used in the pre-processing stage of the proposed approaches. Chapter 4 discusses a proposed algorithm for solving electrode dropout issue. Chapter 5 illustrates alternative representation of image that enhances the object recognition. Chapter 6 discusses four proposed enhancement techniques utilized in a real-time mixed reality simulation. Finally, Chapter 7 outlines the conclusions of the thesis and possible future work extensions.

2

Background

2.1 Overview of the Eye and the Brain

When anyone thinks of vision, the thoughts go directly to the eye. However, the majority of the visual process happens in the brain [6]. The light rays that are focused by the eye pupil pass through the eye lens that redirects all the light to the eye retina. The signals from the retinal photoreceptors (i.e., the rods and the cones) are transmitted to the bipolar cells, until reaching the retinal ganglion cells, where the retina's ganglion cell axons form the optic nerve. The nasal fibers of each eye cross each others at the optic chiasm, continuing to the optic tract with the temporal fibers. Finally, optic radiations connect the lateral geniculate nucleus (LGN) to the primary visual cortex, where the visual information is processed [7].

To further illustrate the visual pathway, Figure 2.1 shows the visual projection pathway from the eye until reaching the brain. The eye captures the light energy that is transmitted to the photoreceptors of the retina which then, converts the light energy to electrical neurons. Then, the electrical neurons are transmitted to the optic nerve where the optic nerve connects the eye to the brain, thus, it is responsible for the transmission of the visual information from the retina to the brain [8]. Moreover, the optic chiasm is formed by the crossings of the optic nerve in the brain. The lateral geniculate nucleus (LGN) of thalamus is the main central connection for the optic nerve to the visual cortex. In addition, the superior colliculus is responsible for the transformation of the sensory input into movement output and the orientation of the eye movements to the objects of interest in the outside world. Finally, the visual cortex is the primary cortical region of the brain that receives, integrates, and processes visual information relayed from the retinas.

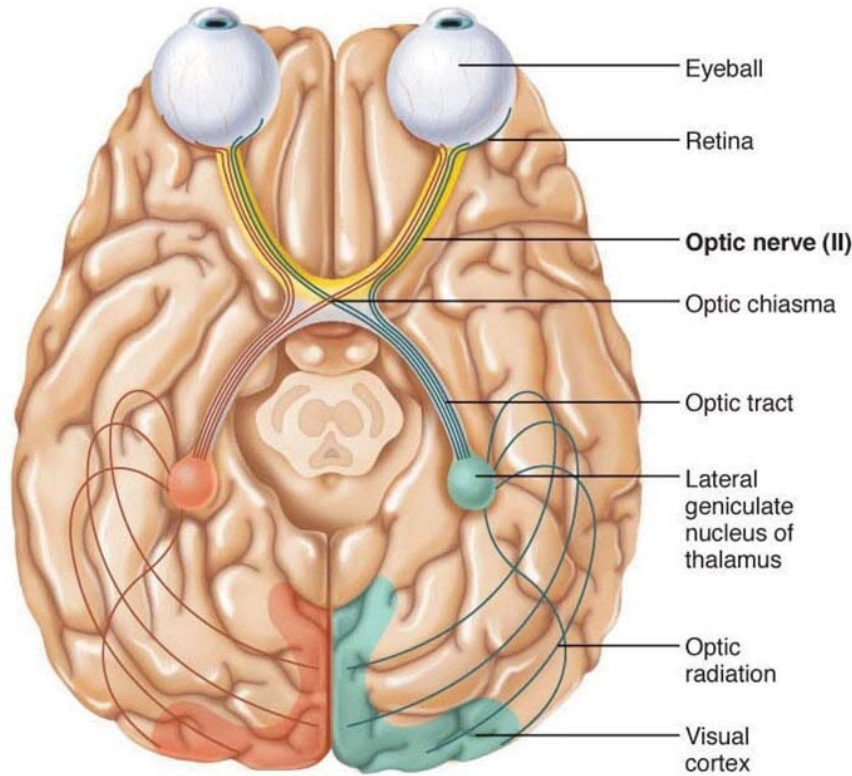


Figure 2.1: The visual projection pathway (Adapted from [8]).

2.2 The Retina

Among all the visual pathway parts, the retina is the most crucial part since it converts the light to a signal (i.e., electrical signal) that the brain understands. The retina is the innermost, light-sensitive layer of tissue of the human eye that creates visual perception as a result of sending nerve impulses via the optic nerve to the visual cortex. To illustrate the anatomy of the human eye retina, Figure 2.2 shows the structure of a typical retina [9]. The outer layer of the retina, the photoreceptors layer, is responsible for receiving the light from an image, converting the light to action potentials (i.e., electrical signals) and sending the visual information to the retinal ganglion cells that is responsible for sending the visual information via the optic nerve to the brain. Concerning the photoreceptors, it consists of rods and cones, where the rods are responsible for vision at low light intensities (i.e., dim light or night), while the cones are responsible for vision at high light intensities (i.e., bright light or day time). The bipolar cells are specialized sensory neurons for the transmission of senses, and in the case of indirect visual information transmission, they transmit the visual information from the photoreceptors to the ganglion cells, which will then send the nerve impulses (i.e., after converting light energy into electric impulses) through the optic nerve to the visual cortex in the brain. A single layer of regular polygonal cells positioned at the retina's outermost layer makes up the retinal pigment epithelium (RPE), where the inner side of the RPE is related to the outer segment of photoreceptor cells. For the bipolar cells, they actually come in two varieties, which may be identified by the way they react to light shining on the centres of their receptive fields. They are known as OFF-centre cells and ON-centre cells [10]. A bipolar

cell is an ON-centre cell if a light stimulus delivered to the centre of its receptive field has an excitatory impact on it, leading it to become depolarized (i.e., less negative). However, a light ray that simply illuminates the surrounding area will have the opposite effect on this type of cell, inhibiting (hyperpolarizing), or making it more negative. Bipolar cells that are OFF-centre cells exhibit the exact opposite behaviour. Light on the field's perimeter has an excitatory (depolarizing) effect, whereas light on the field's centre inhibits (hyperpolarizes) motion [11]. Finally, the nerve fiber layer comprises the axons of the ganglion neurons coursing on the vitreal surface (i.e., glass-like surface) of the retina to the optic disk. The ganglion cell layer comprises retinal ganglion cells and displaced amacrine cells. The inner plexiform layer includes synaptic connections between the axons of bipolar cells and dendrites of ganglion cells. The inner nuclear layer relays the visual signals from the photoreceptors to the ganglion cells, and represent the signals into visual streams. The outer plexiform layer includes dense network of synapses joining dendrites of horizontal cells from the inner nuclear layer with photoreceptor cell inner segments from the outer nuclear layer [12].

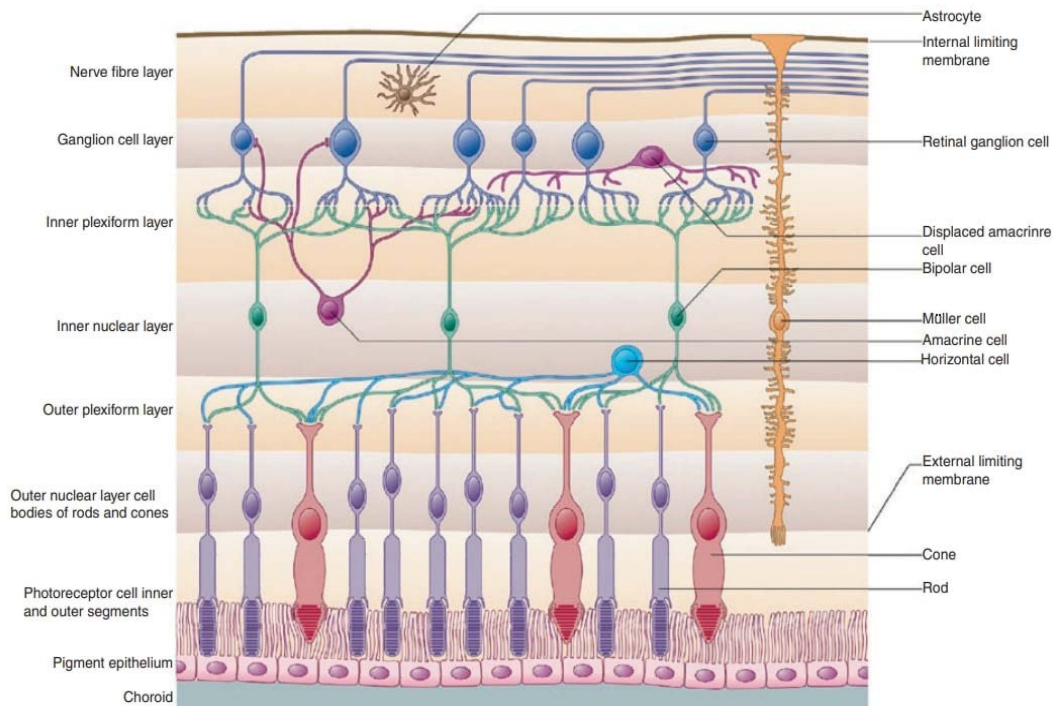


Figure 2.2: Human eye retina anatomy (Adapted from [9]).

2.3 Visual Prosthesis

In the 18th century, Charles Leroy, a French physician and scientist, used two cables, one wrapped around the patient's head just above the eyes and the other around the leg, to discharge static electricity from a Leyden jar, a precursor of modern-day capacitors, into the blind man's body in 1755 [13]. Due to a high fever, the patient, who had been blind for three months, compared the sensation to a flame moving in front of his eyes. For the first time, a rudimentary prosthetic electrical device successfully restored even a flicker of visual perception. Since then, the notion of recovering vision in retinas affected by severe

neurodegenerative disorders has advanced from a far-off dream to a present-day possibility. By electrically stimulating the remaining inner retinal neurons with a device implanted on the retinal surface, damaged photoreceptors can be bypassed. Over the past ten years, the development of such retinal prosthesis has advanced steadily. Recent research on human patients has revealed that long-term low-resolution implants can restore basic vision and pattern recognition. In the 20th century, the first time the human occipital pole was exposed while being electrically stimulated was by German neurosurgeon Förster [14]. In 1929, he discovered that depending on where the stimulation probe was positioned, electrical stimulation caused people to perceive light points known as phosphenes, which are typically described as stars in the sky, clouds, and pinwheels. In a later time during the same century, an electrical engineer, named Robert Greenberg observed a retinal surgeon place a tiny wire into the blind patient’s eye in the operating room in 1991 [15]. The patient reported seeing a pinpoint of light in his completely dark field of vision when the wire touched his retina and provided a tiny shock of electric current. The patient noticed two points of light after the surgeon introduced a second wire. This led to the hypothesis that using diverse light sources will help the blind patients perceive an image. This is what visual prosthesis is all about.

The visual prosthesis, aka bionic eye, is used as a substitute for a damaged part in the visual pathway due to a disease, an infection, an accident or a genetic disorder that caused the loss of vision. The most commonly known reason for loss of vision is due to a genetic disorder where this disease is called Retinitis Pigmentosa or due to an age-related disease called Aged-Macular Degeneration. Retinitis Pigmentosa (RP) is a genetic disease that people are born with, however the disease develops at a later stage in their life, giving hope for partial restoration of vision via visual prostheses. Moreover, the loss of vision due to RP occurs at the periphery [16]. For Aged-Macular Degeneration, it is a medical condition that may result in blurriness or no vision in the center of the visual field. It typically occurs in elder people [17].

The visual interface of the visual prosthesis device, a headset made up of opaque glasses with an integrated video camera, and a “pocket computer,” or a visual processing unit, are the new external components that are being introduced at this point. The “pocket computer” is the VPU that is connected to the headset by a cable. The processing unit is about 4-5 inches long, can be worn around the neck, pocket, or belt, and has a number of control switches that let the user switch between the device’s various perceptual modes (i.e., depending on the device, there are between 3 and 4 different image processing modes, such as edge detection, motion detection, and white-on-black) [18].

A generic illustration of a visual prosthesis device (i.e., Argus II) is shown in Figure 2.3, where a tiny camera is mounted in the bridge of the glasses capturing images and sends them via a wire to video processing unit (VPU). The VPU transforms the images into electrical signals and sends them back to an antenna on the glasses. The signals are then transferred wirelessly to a receiver in the implant, which is the electrode array as shown in Figure 2.3c, enabling the implanted patient to perceive the external environment.

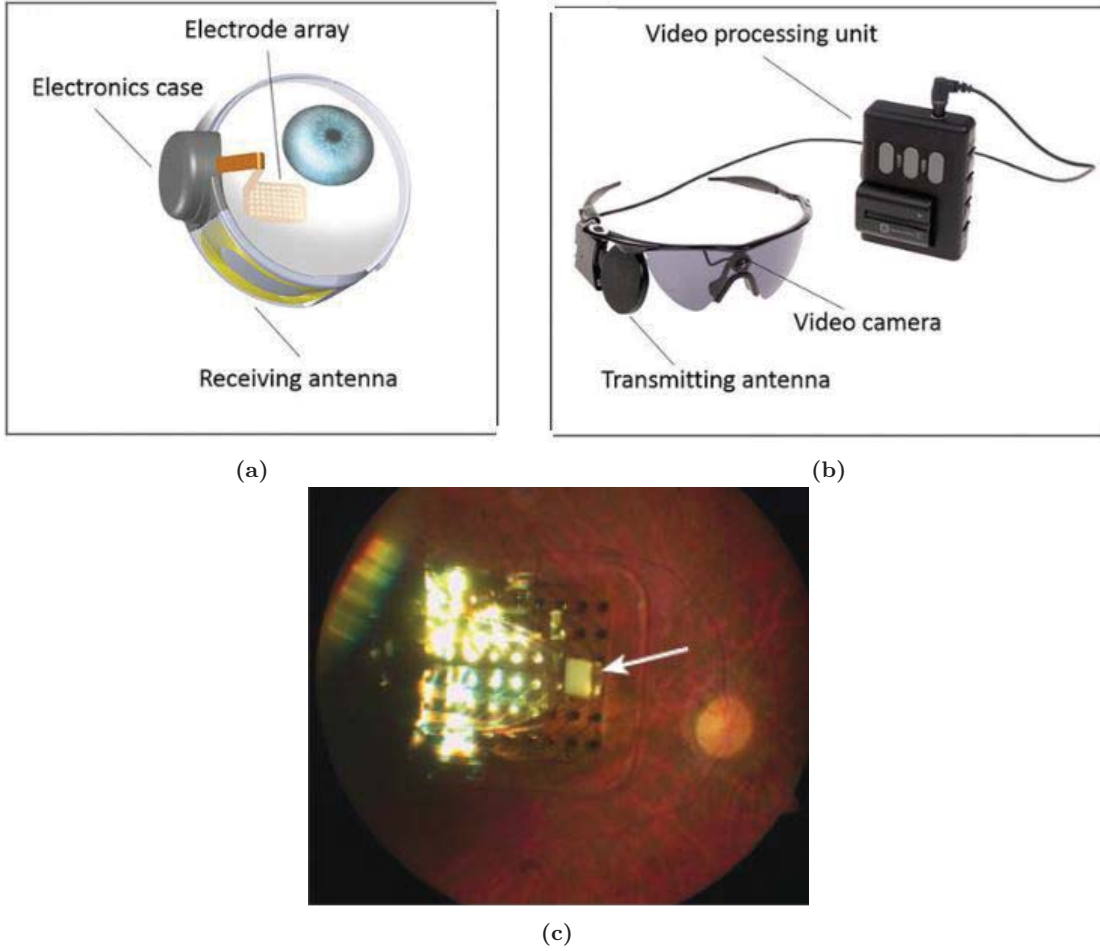


Figure 2.3: Visual prosthesis. (a) Implanted components of the system. (b) External Components of the system. (c) Electrode array implanted in the retina. (Adapted from [19, 20]).

2.3.1 Nature of Stimulation Signals

A visual prosthesis takes setup and stimulation data from an external module, which is subset from the visual prosthesis system, much like any other implantable microstimulating system does [21]. An embedded centralised controller that is in charge of managing the implant functioning and producing stimulation commands/data is part of the implantable microsystem. The system level requirements for embedded controllers, created to be used in all types of visual prostheses, are generally the same. A stimulation back-end generates the proper stimulation pulses and delivers them to the target tissue in response to the stimulation data that the embedded controller has provided. According to the information it receives from the external module, the integrated controller instructs the microstimulator to produce electrical pulses with a range of amplitudes and pulse lengths. This block also regulates other variables including the inter-phase interval and the stimulation period. Additionally, the inbuilt controller manages the timing and arrangement of all electrode stimulations. To induce retinal stimulation by converting the spatiotemporal pattern of light into electrical currents, a single photodiode cannot effectively convert the usual natural retinal irradiance's ($0.001 - 1 \mu W/mm^2$) light intensity into electrical current for brain activation. Based on patient feedback on visual

perception, the amplitude and timing characteristics of the stimulation pulses need to be adjusted. One possible solution for this could be by utilizing an Application Specific Integrated Circuit (ASIC) with a Complementary metal–oxide–semiconductor (CMOS) amplifier circuit within each pixel to amplify the primary photocurrent [22].

Power dissipation in the human eye is one of the key concerns in wireless telemetry [23]. This power loss may cause biological tissues to warm up too quickly, which could cause tissue injury. The quantity of absorbed radio frequency (RF) energy must be assessed when developing wireless interfaces and compared to human safety norms, which are typically defined in terms of specific absorption rate (SAR) in related standards.

2.3.2 Phosphenes

The main aim behind the implementation of the proposed work is to enhance the images retrieved by the camera in the visual prosthetic device so that the implanted patient can see a better quality image with a self-explanatory scene. Thus, the person will be able to live his/her life independently and will be able to build his/her self-confidence that was lost before due to lack of autonomy. The electrodes in the implant correspond to spots of light called phosphenes that are displayed in the visual field [24]. Phosphenes are the essential building elements for using a bionic eye device to provide visually impaired people with meaningful visual information, as shown in Figure 2.4 where the phosphene simulation for a living room is shown. Researchers may be able to change a variety of stimulus parameters in clinical trial applications, such as current amplitude, duration, frequency, etc. Due to the limited wireless transmission bandwidth, the number of perceived gray levels is relatively small which affects the ability of object's recognition [21].

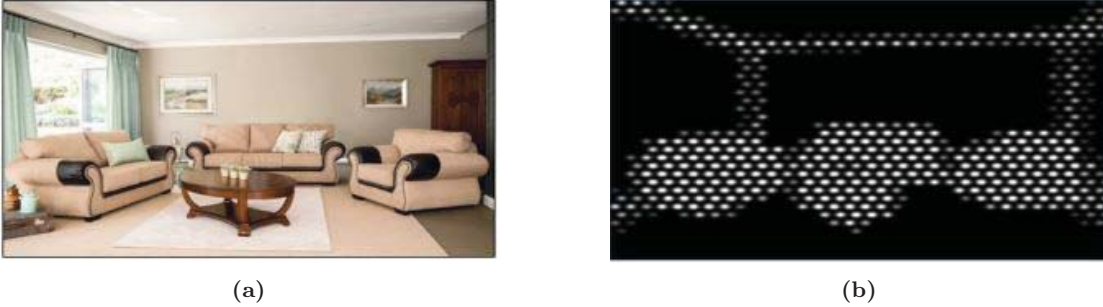


Figure 2.4: Phosphenes sample. (a) Input image. (b) Phosphene simulation. (Adapted from [25]).

2.3.2.1 Phosphene Simulation

To mimic the environment that a typical visual prosthetic user encounters, phosphenes are being simulated such that the actual phosphenes properties are preserved to be as close as possible to the actual phosphenes' perception. In general, phosphene simulation is used by researchers to try to enhance the image perception before clinical trials on real patients [26]. Circular, squared and hexagonal shapes of the phosphenes were performed in literature [27, 28]. Moreover, hexagonal grid and squared grid were the regular grids used in literature [29]. A sample of the squared and hexagonal grid is shown in Figure 2.5 for a cup. The squared grid

is preferable since it gives better visual acuity compared to the hexagonal grid [29]. Moreover, three phosphene simulations were tried in the literature. Ideal phosphene simulation (i.e., round shaped phosphenes with fixed radius), scoreboard phosphene simulation (i.e., round shaped with varying radii based on pixel intensity) and axon map phosphene simulation (i.e., tail-like phosphenes) [18, 30, 31]. For the phosphene shape used in ideal phosphene simulation, circular shape was utilized [28]. For the ideal phosphene simulation, Figure 2.5b shows the ideal phosphene simulation in a squared grid. Whereas Figure 2.5c shows the phosphene simulation of a cup in hexagonal grid. Regarding the scoreboard phosphene simulation, two versions were implemented. The first version is a binary scoreboard phosphene simulation, as shown in Figure 2.6b where a cup is displayed (Figure 2.6a) and the second approach is a grayscale phosphene simulation, as shown in Figure 2.6c where the same cup is displayed. In Figure 2.6b, the size of the rounded phosphenes is controlled by the actual intensity of the pixel and all the phosphenes are given the same white color. Whereas in Figure 2.6c, the phosphenes are given the same color as that of the original pixel intensities and their sizes are controlled by the intensity. So, the higher the intensity value, the larger the radius of the phosphene.

In addition, to simulate the electrode dropout that causes a black spot at the corresponding phosphenes location, Figure 2.7 shows three dropout percentages which are 10%, 20% and 30% [32]. These rates show that the more the dropout rate, the more the ambiguity of the object and the less the ability to recognize the object.

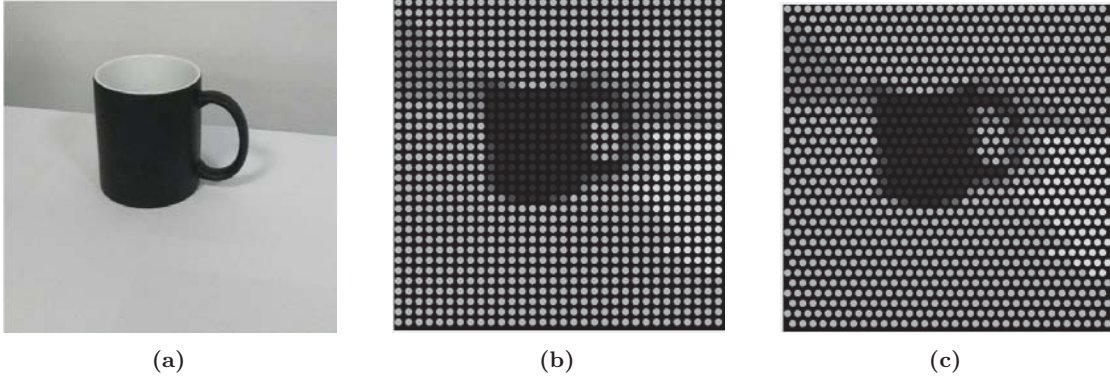


Figure 2.5: Squared and hexagonal grids. (a) Original image. (b) Squared-grid phosphene simulation. (c) Hexagonal-grid phosphene simulation.

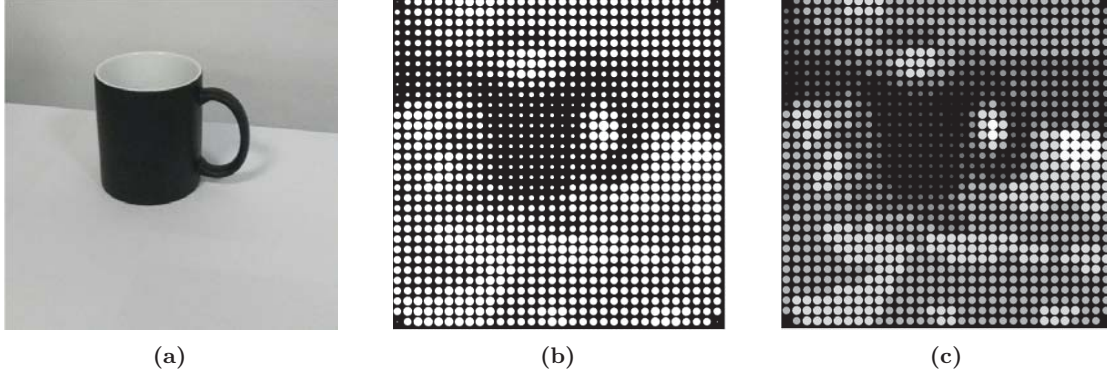


Figure 2.6: Binary and grayscale scoreboard phosphene simulation. (a) Original image. (b) Binary scoreboard phosphene simulation. (c) Grayscale scoreboard phosphene simulation.

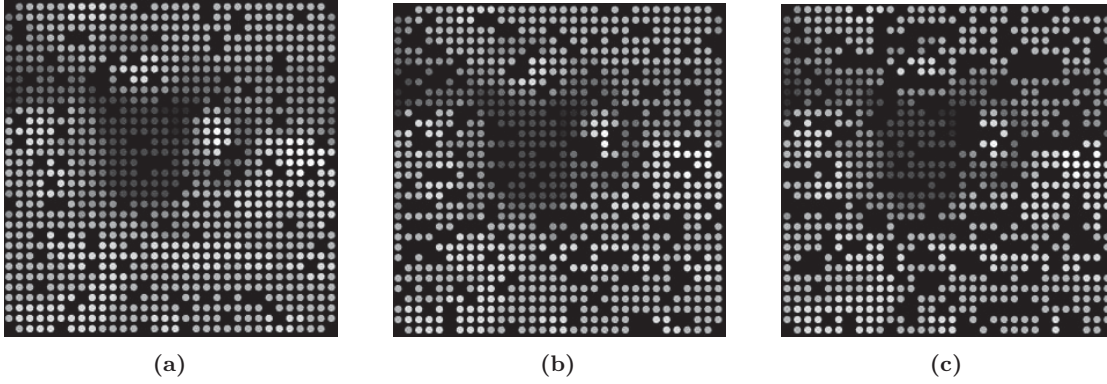


Figure 2.7: Dropout rates of (a) 10%, (b) 20% and (c) 30%, respectively.

2.4 Visual Prostheses Types

2.4.1 Electrode-Based Visual Prostheses

Various electrode-based visual prostheses types are utilized worldwide. The most utilized visual prostheses type so far is retinal prosthesis, since its surgery is the least serious among the other types. The implant is either in the vitreous, close to the inner layers of ganglion cells and nerve fibres (epiretinal prostheses) or between the bipolar cell layer and the retinal pigment epithelium (subretinal prostheses) [33]. Another visual prostheses type that exists between the eye and the brain is the optic nerve prosthesis, where the exterior surface of the optic nerve is surgically covered with an electrode cuff. The system relies on the idea of retinotopic organisation within the optic nerve as it does not pierce the optic nerve sheath [34]. Other types are available where the implant takes place inside the brain. One type is thalamic visual prosthesis, where the implant is placed on the lateral geniculate nucleus (LGN) of the thalamus. This type provides those with diseases affecting the retinal ganglion cells, such as glaucoma, the main cause of blindness for which there is no effective medical cure, with the ability to see again [35]. The second type is visual cortical prosthesis, where the electrodes are placed on the surface of the brain stimulating the neural cells in the visual cortex, enabling the implanted patient to regain partial restoration of vision [36].

2.4.1.1 Comparison of Different Approaches

A comparison between the different visual prosthesis approaches, as illustrated in Table 2.1, reveals different properties for the aforementioned visual prosthesis types.

Table 2.1: Comparison between different visual prosthesis approaches (Adapted from [37]).

Visual Prosthesis					
Point of Comparison	Epiretinal	Subretinal	Optic nerve	Thalamic	Cortical
Population of blind patients concerned	Retinal Degenerations (RP, AMD)	Retinal Degenerations (RP, AMD)	Retinal Degenerations (RP, AMD)	Largest	Largest
Surgical complications	Least serious	Least serious	Less serious	Serious	Serious
Mechanical stability	Difficult	Easy	Easy	Easy	Easy
Visuotopic stimulation	Easy	Easy	Difficult	Difficult	Difficult
Preserves peripheral visual processing	Yes	Yes	No	No	No
Fully implantable	No	Yes	No	No	No
Processing flexibility	Yes	Possible	Yes	Yes	Yes
Large scale integration	Difficult	Easy	Difficult	Difficult	Difficult

2.4.2 Optogenetic-Based Visual Prostheses

A new trend in visual prostheses is the use of optogenetics which is the study of genetic photosensitization of neural tissue [38]. Optogenetic visual prosthesis holds significant potential for novel strategies. After the loss of the rod and cone photoreceptors, the remaining light-insensitive retinal neurons undergo genetic conversion to photosensitive cells, which results in the development of retinal light sensitivity. Since the eye is transparent, it may be less invasive, more affordable, and more effective than current methods to photosensitize the remaining neuronal layers of the eye and illuminate from the outside. In optogenetics, nerve cells are genetically altered to increase their sensitivity to specific light wavelengths. Optogenetic retinal prostheses differ significantly from conventional ones (electrode-based visual prostheses). In order to make specific cells in a target region of the remaining retina light sensitive, a viral vector would need to be injected clinically. The patient could then put on a customized virtual reality (VR) or augmented reality (AR) headset that can transmit the proper light pulses since the eye is transparent. There is no need for intrusive surgery. However, implanted

optoelectronics would be required for brain level visual prostheses. Optoelectronics do not need to be as complexly engineered or as biocompatible as their implantable electronic counterparts. Therefore, it should be easier to design and less expensive to use optogenetic prostheses. Numerous issues with durability and cellular targeting could also be solved. It should be easy to achieve high resolution by simply raising the resolution of a high brightness display and creating the necessary optics.

2.5 Visual Prostheses User Experience and Challenges

Before being able to use visual prostheses, implanted patients have to learn how to use the prosthetic vision device at the very first glance following the implantation of the implant [18]. Recipients are told about the components of the device and given usage instructions before the camera starts recording. Following the device implantation, the user is instructed on the physical abilities required to utilize the device, such as head and eye alignment with the camera and head movements for scanning. They have to be conscious of and in alignment with their head, camera, and eyes in order to utilize the signal to orient themselves in space as if the camera is their new eye.

In case of retinal prostheses, which is the most common visual prosthesis type, the user is then instructed to try focusing the camera on any desired object in space. Since the camera is not pointed at their eye or pupil, but rather at a distance of a few inches, at the centre of their brow ridge, they have to learn to adjust all movements and estimations of objects in space by those few inches (above the nose, between the two eyes). The instructors typically give the trainees instructions to draw a line in space with their index finger from the camera to the item in order to understand the discrepancy. Then, the recipients are given instructions on how to scan the environment using the head movement to perceive the space and the objects around them. In order to maximize the recipients' perceptual range, they are advised to start making large scanning motions when entering unfamiliar situations. They are then instructed to make increasingly little movements, which refresh the image as they zoom in on an object or particular aspects of interest. The recipient must scan the surroundings, reassembling their fragmentary views in their minds since the visual area covered by the implant is rather small—no more than 20 degrees (approximately the breadth of two hands, outstretched). The camera is turned on two weeks after the first activation and system installation. It is a moment that is frequently met with great excitement since the recipients are informed that this is when they will start to regain a sort of functional eyesight. The recipients stated that things got significantly more chaotic and “noisy” when the camera is active and all of the electrodes can be activated simultaneously.

The electrical device parameters must be customized for the patient after a post-operative period of recovery [39]. Although retinal implants may involve tens or even hundreds of electrodes, the density and connectivity of the remaining healthy inner retinal neurons will have a significant impact on the effectiveness of stimulation and the quality of visual perception. The length of blindness, the condition of the retina, the placement during surgery, the length of the axial globe, and the obligation of the device to the globe's curvature can differ amongst recipients. It is necessary to remember that the perception that implantees have, may not

always correspond to real-world vision or synthetic visuals. Due to the wide range of retinal cell types that are stimulated and the unintentional activation of axon fibres, the experience is frequently spatially distorted and temporally complex. Clinical psychophysics' practical aim is to link stimulus parameters to patient perception in order to provide each recipient with the best device parameters. The threshold charge at which a visual percept is elicited is the most fundamental parameter for a retinal implant, and it varies across electrodes and implanted patients.

Patients reported a range of experiences across months to years, outside of the treatment centres and clinical trial testing rooms [18]. The device may operate past its anticipated lifespan or it may fail early. Some users claim that after using it to gain a sensation of the familiar objects in their own home, they believe they have no further use for it. After 1-2 years, users claim to have reached a dead end and are disillusioned by the technology. Due to this, many recipients simply quit using their device after a while. So, a possible solution could be through providing image enhancement techniques that will motivate the visual prostheses users to remain using the device [40].

Although the introduction of visual prostheses gives hope for blind patients to partially restore their vision, there are some challenges that a typical visual prosthetic user suffers from. Implanted patients reported that their perceptual experiences were fundamentally and significantly different from what they would perceive naturally [18]. Due to the limited number of electrodes in the implant, low spatial and radiometric resolutions in a typical visual prostheses system are perceived affecting the quality of the perceived image [41]. This results in the loss of the details preservation that hinders the ability of any implanted patient to correctly recognize or localize an arbitrary object. Some system-level limitations and restrictions, such as wireless transmission bandwidth and implanted module processing power, affect the number of gray levels perceived [42]. Thus, the quality of the perception is also reduced. Furthermore, over time, electrode malfunctions that cause dropouts at the relevant place in the visual field are possible [43]. Figure 2.8 demonstrates a simulation for the effect of having dropout added (i.e., DO as shown in the figure) and no dropout added (i.e., ND as shown in the figure) in addition to having a limited number of electrodes, resulting in loss of the crucial details of the face.

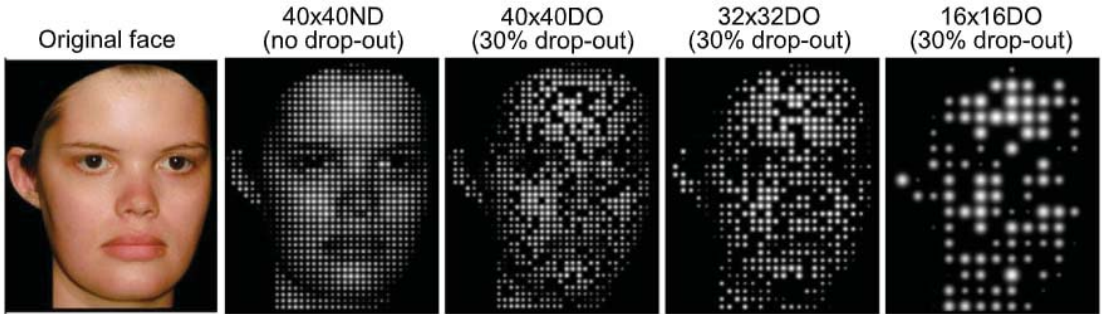


Figure 2.8: Face sample in simulated prosthetic vision with no dropout and with dropout added (Adapted from [44]).

2.6 Image Processing Techniques

2.6.1 Techniques Used

2.6.1.1 Histogram of Oriented Gradients

To extract features from image data, feature descriptors such as Histogram of Oriented Gradients (HOG) are frequently utilized [45]. It is commonly used for object detection in computer vision tasks. The HOG descriptor emphasizes an object's structure or shape. This is accomplished by extracting the edges' gradients and edges' orientations. These orientations are computed in discrete, "localized," areas. This implies that the entire image is divided into smaller sections, and that the gradients and orientations are calculated for each region. Finally, the HOG would create a distinct histogram for each of these regions. The gradient magnitudes and orientations of the pixel values are used to build the histograms. The gradient magnitude of a pixel can be calculated as,

$$\text{Gradient Magnitude} = \sqrt{(G_x)^2 + (G_y)^2} \quad (2.1)$$

where G_x and G_y are the gradients in the x and y directions, respectively. On the other hand, the gradient orientation of a pixel can be calculated as

$$\phi = \tan^{-1} \frac{G_y}{G_x} \quad (2.2)$$

HOG can be utilized to find the best match of a query image in a set of images. This could be done by getting the descriptor for each of the images in the dataset and the descriptor for the query image. Then, calculate the Euclidean distance between each of the descriptors of the dataset images and the descriptor of the query image to get the minimum distance. Thus, the minimum distance indicates the best matching image from the dataset compared to the query image. The Euclidean distance is computed as,

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^N (v_1[i] - v_2[i])^2} \quad (2.3)$$

where N is the number of elements in the descriptor, v_1 is the descriptor vector of the query image and v_2 is the descriptor vector of the image from the dataset.

2.6.1.2 Otsu Thresholding

One typical image processing operation is to change a grayscale image to monochrome [46]. Nobuyuki Otsu is the creator of the Otsu method, one of numerous binarization methods. In Otsu's thresholding method, each conceivable threshold value is iterated over in order to determine the spread for the pixel levels on either side of the threshold, or the pixels that are either in the foreground or background. Finding the threshold value at which the total of foreground and background spreads is at its lowest is the goal.

To illustrate the Otsu thresholding mathematically, a threshold is needed to divide the pixels of an image to foreground and background pixels. All possible thresholds, from 0 till 255, should be examined. Equations 2.4, 2.5 and 2.6 calculate the foreground and background variances. The weight W of either the foreground pixels or the background pixels is computed as,

$$W = \frac{\sum_{i=0}^{255} g[i]}{M \times N} \quad (2.4)$$

where g is the total number of pixels having a certain intensity from the range 0 till 255, M and N are the width and the height of the image, respectively. The mean μ of either the foreground pixels or the background pixels is computed as,

$$\mu = \frac{\sum_{i=0}^{255} (I[i] \times g[i])}{Q} \quad (2.5)$$

where I is the pixel intensity, g is the total number of pixels having the same intensity and Q is the total number of pixels for all the intensities consumed. The variance σ^2 of either the foreground pixels or the background pixels is calculated as,

$$\sigma^2 = \frac{\sum_{i=0}^{255} ((I[i] - \mu)^2 \times g[i])}{Q} \quad (2.6)$$

where I is the pixel intensity, μ is the mean value of either the foreground or the background, g is the total number of pixels having the same intensity and Q is the total number of pixels for all the intensities consumed.

Then, to calculate the within-class variance σ_W^2 , it can be calculate as

$$\sigma_W^2 = W_b \sigma_b^2 + W_f \sigma_f^2 \quad (2.7)$$

where W_b is the weight of the background, σ_b^2 is the variance of the background, W_f is the weight of the foreground and σ_f^2 is the variance of the foreground.

To have a faster calculation of the desired threshold, between-class variance can be used instead of within-class variance to eliminate the utilization of the complex mathematical operations used in Equation 2.7 (i.e., the complexity of the equation that calculates σ^2 , presented in Equation 2.6). The between-class variance σ_B^2 is computed as,

$$\sigma_B^2 = W_b W_f (\mu_b - \mu_f)^2 \quad (2.8)$$

where W_b is the weight of the background, W_f is the weight of the foreground, μ_b is the mean value of the background and μ_f is the mean value of the foreground.

Finally, the best threshold that will be chosen at the end is the one with the maximum between-class variance (σ_B^2) which is at the same time has the minimum within-class variance (σ_W^2).

2.6.1.3 Dilation

One of the fundamental operations in mathematical morphology is dilation [47]. It is usually represented by “ \oplus ”. Dilation was originally developed for binary images, however, it can be used in grayscale images. A structuring element is typically used by the dilation process to probe and expand the shapes present in the input image. Pixels are added to object borders through dilation. The maximum value of all the pixels in the surrounding area makes up the output pixel’s value. A pixel is set to 1 if any of its adjacent pixels also have a value of 1. Figure 2.9 shows an image where dilation is applied on using a “cross” structuring element. The black pixels represent the points added by the dilation. Equation 2.9 shows the binary dilation operation which is widely used in image processing.

$$A \oplus B = \bigcup_{b \in B} A_b \quad (2.9)$$

where A is the binary image, B is the structuring element and A_b is the translation of A by b .

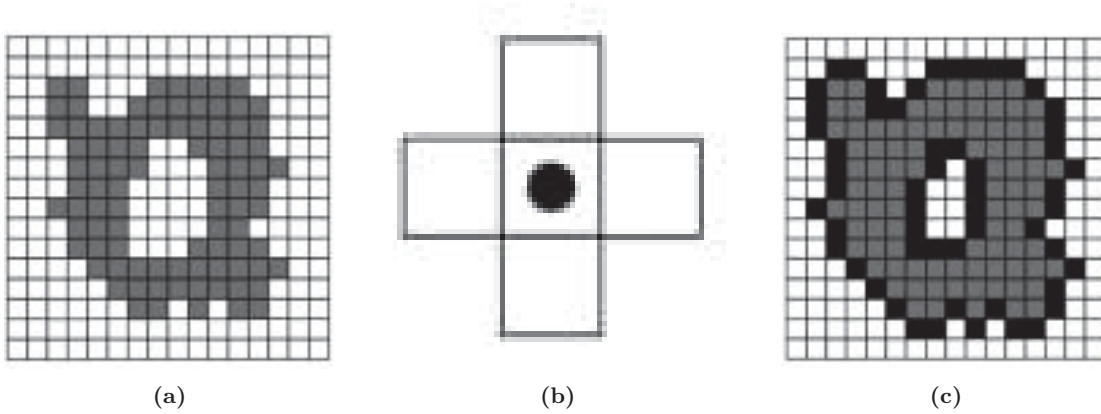


Figure 2.9: Dilation applied on input image. (a) Input image. (b) Cross structuring element. (c) Dilated image. (Adapted from [48]).

2.6.1.4 Skeletonization

Skeletonization, aka medial axis transform (MAT), is a morphological operation that provides an effective way to describe and evaluate the topology of objects [49]. Skeletonization provides region-based shape features. It is a typical pre-processing procedure in pattern recognition and raster to vector conversion. There are two primary methods for creating the skeleton. The first is to employ some sort of morphological thinning that gradually removes pixels from the boundary (while maintaining the end points of line segments) until further thinning is impossible, at which point what is left roughly resembles the skeleton. The alternate approach is to compute the image’s distance transform first. The skeleton then lies along the singularities in the distance transform, which are creases or curvature discontinuities. The MAT may be calculated using the latter method better because it is the same as the distance transform but with all points outside of the skeleton suppressed to zero. The skeletonization for a human is shown in Figure 2.10, where the topology of the human is preserved.

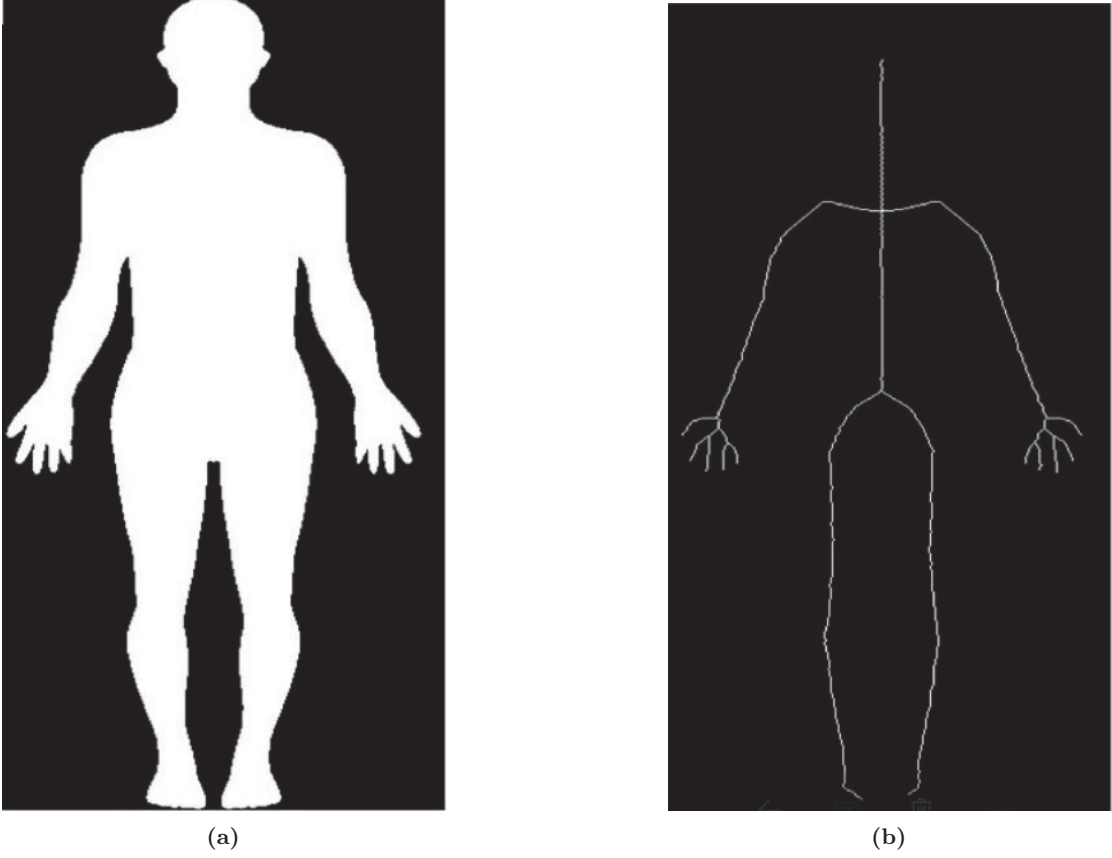


Figure 2.10: Skeletonization sample. (a) Original image. (b) Skeletonized image. (Adapted from [50]).

The skeleton for a continuous binary image can be computed as,

$$S(X) = \bigcup_{\rho > 0} \bigcap_{\mu > 0} [(X \ominus \rho B) - (X \ominus \rho B) \circ \mu \bar{B}] \quad (2.10)$$

where \ominus and \circ are the morphological erosion and opening, respectively, ρB is an open ball (i.e., a solid sphere that excludes the bounding points that outlines the sphere) of radius ρ , B is the structuring element, and \bar{B} is the closure of B . Whereas the skeleton for a discrete binary image is the skeleton subsets $S_n(X)$ that can be calculated as,

$$S_n(X) = (X \ominus nB) - (X \ominus nB) \circ B \quad (2.11)$$

where n is from 0 till N as N is the number of skeleton subsets. In other words, n is the size of the structuring element. In this case, $S(X)$ is the union of $S_n(X)$.

2.6.1.5 Connected Components Labelling

In order to ascertain the connectivity of “blob”-like regions in a binary image, a technique known as connected component labeling (CCL), blob extraction, or region labeling, is an algorithmic application of graph theory [51]. Although colour images and data with larger dimensionality can also be analyzed, linked-component labelling is employed in computer vision to locate connected parts in binary digital images. A common technique that is utilized

2. Background

in the resulting binary image after thresholding is blob extraction, which can also be utilized in grayscale and colored images. Two types of connectivity exist which are 4-connectivity and 8-connectivity. To illustrate the 4-connectivity, let a pixel p at the coordinate (x, y) where x lies between 0 till $N-1$ (i.e., N is the width of the binary image) and y lies between 0 till $M-1$ (i.e., M is the height of the binary image), be denoted as $p(x, y)$, then, its four neighbouring pixels will be $p(x-1, y)$, $p(x, y-1)$, $p(x, y+1)$ and $p(x+1, y)$. Similarly, in 8-connectivity, the neighbouring pixels for a pixel p are the same as that in the 4-connectivity in addition to the diagonal pixels which are $p(x-1, y-1)$, $p(x+1, y-1)$, $p(x-1, y+1)$ and $p(x+1, y+1)$, as shown in Figure 2.11. Thus, object pixels connected to each other, are given the same label since they belong to the same object. So, in order to ensure that pixels belonging to the same object will be connected together, as shown in Figure 2.12, it is recommended to utilize 8-connectivity rather than 4-connectivity.

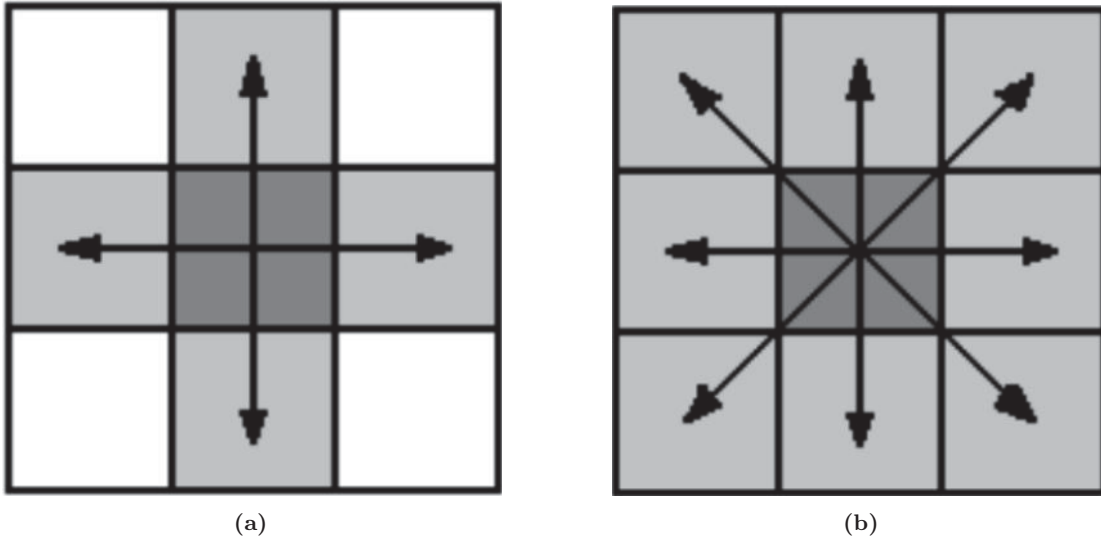


Figure 2.11: Connected components labelling. (a) 4-connectivity. (b) S8-connectivity. (Adapted from [51]).

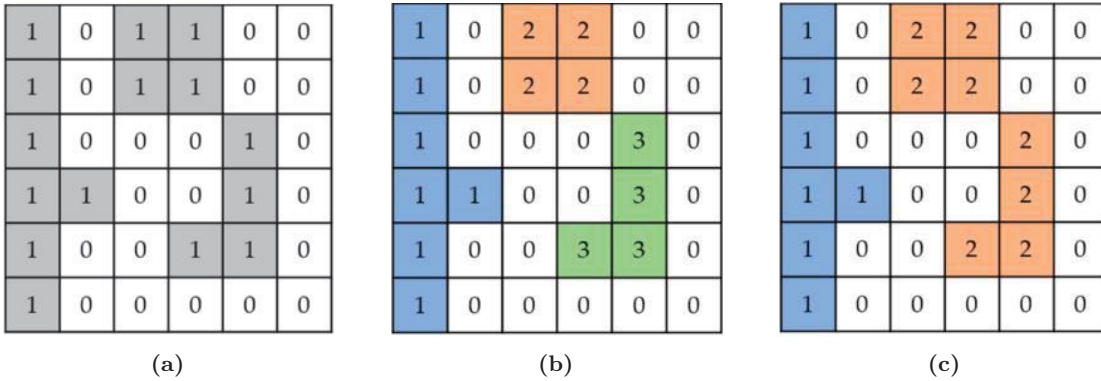


Figure 2.12: Connected components labelling algorithm. (a) Binary image. (b) Three objects are detected by 4-connectivity component labelling. (c) Two objects are detected by 8-connectivity component labelling. (Adapted from [52]).

2.6.1.6 Features from Accelerated Segment Test

Features from Accelerated Segment Test (FAST) is an algorithm that Rosten and Drummond first suggested for locating interesting areas in an image [53]. A pixel in an image that has a distinct position and can be reliably recognized is considered to be an interest point. Interest spots should preferably be repeated across photos and have a high level of local information content. Application areas for interest point detection include object recognition, tracking, and image matching. Interest point or corner detection is not a brand-new concept in literature; the terms are often used interchangeably. There are a number of well-known algorithms, including the smallest univalue segment assimilating nucleus (SUSAN) corner detector, the Harris & Stephens corner detection technique, and the Moravec corner detection algorithm. Creating an interest point detector for use in real-time frame rate applications like simultaneous localization and mapping (SLAM) on a mobile robot, which have constrained processing resources, was the goal of the work on the FAST method.

The following illustrates the machine learning approach for the FAST algorithm. First, a set of images for training is selected. Second, the FAST algorithm is run on each image to find the focal points by analysing the 16 pixels in the circle (i.e., Bresenham circle of radius 3) one at a time. Third, the method keeps the 16 pixels around each pixel “ p ” in a vector form, as shown in Figure 2.13.

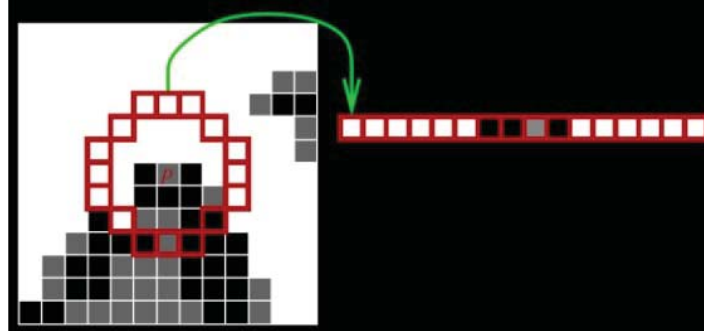


Figure 2.13: The 16 values surrounding pixel p stored in vector form (Adapted from [54]).

Fourth, this is repeated for each and every pixel in every image. This is the vector P that includes all of the training data. Fifth, three states are possible for each value. Consider one of the 16 pixels, x , in the vector, as shown in Equation 2.12, is greater than p , less than p , or similar to p .

$$S_{p \rightarrow x} = \begin{cases} d, & I_{p \rightarrow x} \leq I_p - t \\ s, & I_p - t < I_{p \rightarrow x} < I_p + t \\ b, & I_p + t \leq I_{p \rightarrow x} \end{cases} \quad (2.12)$$

where $S_{p \rightarrow x}$ is the state, d denotes darker than p , s denotes similar to p , b denotes brighter than p , $I_{p \rightarrow x}$ is the intensity of the pixel x and t is a threshold. Sixth, the entire vector P will be partitioned into three subsets, P_d , P_s , and P_b , depending on the states. Seventh, the variable K_p is created, which will be true if point p is an interest point (i.e., a pixel that can be reliably identified and has a well defined position) and false otherwise. Eighth, using the

variable K_p , use the ID3 algorithm (decision tree classifier) to query each subset a question regarding the true class. Ninth, entropy minimization is the foundation of the ID3 algorithm. Query as few questions as necessary to identify the true class (whether it is an interest point or not) from the 16 pixels. Or in other words, pick the pixel x that contains the most details about the pixel p . Mathematically, the entropy for the set P can be expressed as,

$$H(P) = (c + \bar{c}) \log_2(c + \bar{c}) - c \log_2 c - \bar{c} \log_2 \bar{c} \quad (2.13)$$

where $c = |p|K_p \text{ is true}|$ denotes the number of corners detected and $\bar{c} = |p|K_p \text{ is false}|$ denotes the number of non-corners detected. Tenth, this entropy minimization is applied iteratively to each of the three subsets. When a subset's entropy is zero, the procedure should come to an end. Finally, it is possible to use the decision tree's learned querying order to speed up detection in other images.

2.6.1.7 Contrast Limited Adaptive Histogram Equalization

A computer image processing method called adaptive histogram equalization (AHE) is used to enhance contrast in images [55]. The adaptive approach is different from traditional histogram equalization in that it computes many histograms, each corresponding to a different area of the image, and then uses them to disperse the image's brightness values. Therefore, it is appropriate for strengthening the definition of edges in each area of an image as well as the local contrast. AHE has a propensity to exaggerate noise in relatively homogeneous areas of an image, though. This is avoided by contrast limited adaptive histogram equalization (CLAHE), a type of adaptive histogram equalization, by restricting the amplification. The slope of the transformation function determines the contrast amplification near a given pixel value in CLAHE. This is proportional to the histogram's value at that pixel value and the slope of the neighbourhood cumulative distribution function (CDF). Before calculating the CDF, CLAHE clips the histogram at a predetermined value to reduce the amplification. As a result, the transformation function's slope is constrained. The so-called clip limit, or value at which the histogram is clipped, is a function of the histogram's normalization and, consequently, of the size of the neighbourhood region. The resulting amplification is typically limited to between 3 and 4. It is preferable to disperse the portion of the histogram that exceeds the clip limit equally among all of the histogram bins rather than discarding it. As a result of the redistribution, some bins will once more cross the clip limit, resulting in an effective clip limit that is higher than the legal limit and whose precise value depends on the image. If this is undesirable, the redistribution process can be repeated until the excess is negligible. CLAHE uses Rayleigh distribution function where the function can be calculated as,

$$\text{Rayleigh } g = g_{min} + \left[2(\alpha^2) \ln \left(\frac{1}{1 - P(f)} \right) \right]^{0.5} \quad (2.14)$$

where g_{min} is a minimum pixel value, $P(f)$ is a cumulative probability distribution and α^2 is a non-negative real scalar specifying a distribution parameter.

2.6.1.8 Median Filter

Due to its effective performance for some particular noise types, such as “Gaussian,” “random,” and “salt and pepper” sounds, the median filter is one of the well-known order-statistic filters [56]. The median filter replaces a $M \times M$ neighborhood’s centre pixel with the median value of the accompanying window. By implementing this concept, a median filter can get rid of these noise issues. A median filter is a nonlinear filter that computes each output sample as the median value of the input samples inside the window; the outcome is the middle value following the sorting of the input values. The main purpose of utilizing median filter is to preserve edges present in digital images. To perform median filter to an image, zero padding is done (i.e., adding two extra rows of zeros (one at the top and the other at the bottom) and two extra columns of zeros (one at the left position and the other one at the right position) so that when the kernel (i.e., the median filter) is applied, overlapping occurs. Then, the kernel iterates over the padded image, making its center aligned with each of the original pixels. Consequently, sorting is applied for the overlapping $M \times M$ values, where $M \times M$ is the number of pixels in the median filter. The median value for a certain pixel, in case of an even-number of elements, can be calculated as,

$$\text{Median Value} = \frac{a\left(\frac{N}{2}\right) + a\left(\frac{N}{2} + 1\right)}{2} \quad (2.15)$$

where a is the sorted array and N is the number of elements in the array. Whereas, if N is odd-number of elements then, median value is calculated as,

$$\text{Median Value} = \frac{a\left(\frac{N+1}{2}\right)}{2} \quad (2.16)$$

where a is the sorted array and N is the number of elements in the array.

2.6.1.9 Wiener Filter

Assuming known stationary signal-noise spectra and additive noise, the Wiener filter is a filter used in signal processing to provide an estimate of a desired or target random process through linear time-invariant (LTI) filtering of an observed noisy process [57]. The mean square error between the intended process and the estimated random process is reduced by the Wiener filter. By employing a related signal as an input and filtering that known signal to obtain the estimate as an output, the Wiener filter aims to compute a statistical estimate of an unknown signal. For instance, the known signal might be made up of a potentially valuable unknown signal that has been tampered with by additive noise. By removing the noise from the distorted signal, the Wiener filter can estimate the underlying signal of interest. Wiener filters are characterized by the following. It assumes that the stationary linear stochastic processes that make up the signal and (additive) noise have known spectral properties or known auto- and cross-correlations. The filter needs to be causal and physically realizable (i.e., this requirement can be dropped, resulting in a non-causal solution). The performance is measured using minimum mean-square error (MMSE). Typically, Wiener filters are used in the frequency domain. The Wiener filter is defined as,

$$G(u, v) = \frac{H^*(u, v)P_s(u, v)}{|H(u, v)|^2P_s(u, v) + P_n(u, v)} \quad (2.17)$$

where $G(u, v)$ is the Wiener filter, $H^*(u, v)$ is the conjugate of the Fourier transform of the point-spread function (PSF), $P_s(u, v)$ is the signal process's power spectrum as determined by the signal auto-correlation's Fourier transform and $P_n(u, v)$ is the noise process's power spectrum, which may be calculated by applying the noise auto-correlation's Fourier transform.

2.6.2 Related Work

Various image processing and computer vision techniques were used to solve the object recognition problem arising in the visual prostheses due to the low-resolution environment. Patients who are blind or visually impaired need to be able to recognize familiar faces, which is possible with a retinal prosthesis [58]. However, the multichannel electrode arrays utilised in today's visual prosthesis have difficulties when it comes to delivering facial images with a resolution sufficient to detect facial features like the eyes and nose. The work of Chang et al. confirm the viability of recognizing known faces with low-resolution prosthetic vision and suggests a technique for edge augmentation to offer more high-quality visual data. By using the Sobel edge detector, a contrast-enhanced image and an edge image are created, which were then blocked individually by averaging. Then, a pixelized image, that mimicked an array of phosphenes by subtracting the blocked edge image from the blocked contrast-enhanced image, is created. Each gray value of the edge images was given a weight of 50% (mode 2), 75% (mode 3), and 100% (mode 4) before subtraction. When using mode 1, the facial image was simply blocked and pixelized. In terms of identification index, which takes into account both accuracy and correct response time, mode 3 was the most successful identification at every resolution. It was also discovered that even with low-resolution prosthetic vision, individuals were able to recognize a distinguishing face more quickly and accurately than the other facial images. Even in extremely low-resolution photos, every individual could recognize recognizable faces. With mode 4, an accuracy of 87.62% was the best recognition accuracy achieved with the same grid size.

Blind individuals who have electronic visual prostheses, or "bionic eyes," implanted are likely to experience some coarse visual impressions [59]. It is anticipated that the first quality of artificially induced vision will be quite poor. Boyle et al. investigate image processing methods that enhance perception for people who use visual prostheses. To replicate artificially induced vision anticipated from new electronic visual prosthesis designs, visual perception studies were conducted with normally seeing observers. Images that were given to subjects as low-resolution 25×25 binary images had various region-of-interest (ROI) processing techniques applied to them. To assess their viability for use in autonomously managing a zoom-type function for visual prostheses, a number of additional processing techniques were examined. The results of the study demonstrate that ROI processing, when applied in a zoom application, enhances scene understanding for low-quality photos, where 95% confidence interval was achieved.

Zhao et al. examined the effects of two types of image processing techniques, two common shapes of pixels (square and circular), and six resolutions (8×8 , 16×16 , 24×24 , 32×32 , 48×48 and 64×64) in order to identify the key factors in common object and scene

images recognition and optimize the recognition accuracy on low resolution images using image processing strategies [27]. The results indicated that the number of pixels is directly proportional to the mean recognition accuracy. The range from 16×16 to 24×24 pixels was the recognition threshold for objects, whereas it was between 32×32 and 48×48 pixels for simple scenes. Different image modes have a significant impact on recognition accuracy close to the threshold of recognition. The best results for recognition were obtained with images that had “threshold pixel number and binarization-circular points”, giving an accuracy that is exceeding 60%.

To enhance the perception of the dynamic scenes from everyday life, Wang et al. apply two image-processing techniques based on a novel background reduction method [60]. In comparison to techniques that directly combined pixels to reduce resolution, psychophysical data revealed that background reduction or background reduction with foreground enhancement boosted response accuracy. The best performance and highest recognition accuracy were obtained using a background reduction/foreground augmentation method that included more grayscale information, where an accuracy of 80% was achieved when utilizing a resolution of 32×32 . Based on these findings, image-processing components for a visual prosthesis could be developed further to help implanted patients, avoid danger and achieve independent movement in daily life.

Guo et al. suggest two image processing methods which are based on a salient object recognition method [61]. The two processing techniques allow the prosthetic implants to concentrate on the target and block out distracting background noise. Psychophysical studies demonstrate the beneficial effects of methods like foreground zooming with background clutter removal and foreground edge detection with background reduction on the task of object recognition in simulated prosthetic vision. The two processing methodologies considerably increase the recognition accuracy of objects by utilizing edge detection and zooming techniques, where an accuracy of 90% was achieved. Therefore, the blind can benefit from using the visual prosthesis to enhance their object recognition skills. The outcomes offer practical options for the continued development of visual prostheses.

2.7 Deep Learning Techniques

2.7.1 Techniques Used

2.7.1.1 Convolutional Neural Network

Based on how the human visual cortex is organized, Convolutional Neural Network (CNN) is one type of deep learning models for processing data with a grid pattern, such as images [62]. Additionally, it is intended to automatically and adaptively learn spatial feature hierarchies, from basic to complex patterns. Convolution, pooling, and fully linked layers are the three types of layers (or “building blocks”) that make up a standard CNN. Feature extraction is carried out by the first two layers—convolution and pooling—while the third layer—a fully connected layer—maps the extracted features into the final output, such as classification. A crucial part of CNN is played by the convolution layer, which is made up of a stack of mathematical operations, including convolution, a specific kind of linear operation. Since a feature may appear anywhere

2. Background

in a digital image, the pixel values are stored in a two-dimensional (2D) grid, and a small grid of parameters known as the kernel, an optimizable feature extractor, that is applied at each image position, making CNNs extremely effective for image processing. Extracted features may gradually and hierarchically become more sophisticated as one layer feeds its output into the following layer. Training is the process of minimizing the difference between outputs and ground truth labels using an optimization technique called backpropagation and gradient descent, among others. Figure 2.14 shows an overview of a sample CNN architecture where a CNN is composed of a stacking of several building blocks: convolution layers, pooling layers (e.g., max pooling), and fully connected (FC) layers. A model's performance under particular kernels and weights is calculated with a loss function through forward propagation on a training dataset, and learnable parameters, i.e., kernels and weights, are updated according to the loss value through backpropagation with gradient descent optimization algorithm. A nonlinear activation function is then applied to the results of a linear process, such as convolution. Although smooth nonlinear functions like the sigmoid or hyperbolic tangent (tanh) function, as shown in Figure 2.15, have been employed in the past because they are mathematical representations of the behaviour of biological neurons, the rectified linear unit (ReLU) is currently the most widely utilized nonlinear activation function [63].



Figure 2.14: An overview of the training procedure and a convolutional neural network (CNN) architecture (Adapted from [62]).

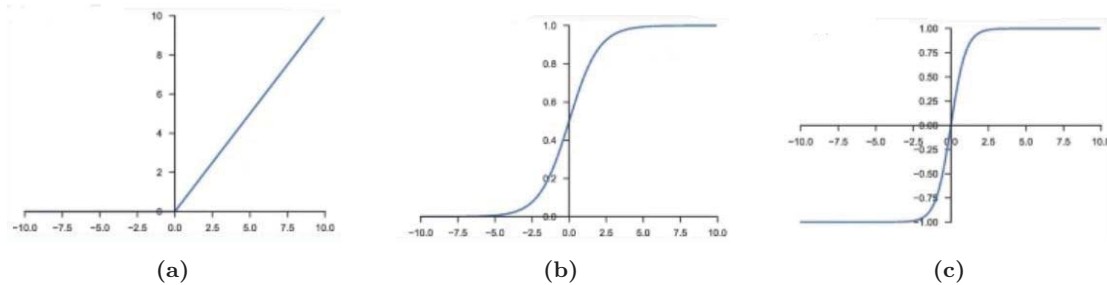


Figure 2.15: Common activation methods for neural networks. (a) rectified linear unit (ReLU). (b) sigmoid. (c) hyperbolic tangent (tanh). (Adapted from [62]).

A model is said to be overfit when it learns statistical regularities unique to the training set, memorizing the irrelevant noise rather than the signal, and then performs poorly on a new dataset after that. One of the fundamental issues with machine learning is that an overfitted model cannot be applied to data that has never been seen before. In this regard, a

test set is crucial to conduct an accurate performance evaluation of machine learning models. Monitoring the loss and accuracy on the training and validation sets is a regular check for spotting overfitting to the training data. The model has probably been overfit to the training data if it performs better on the training set than the validation set. Several strategies have been proposed to reduce overfitting. Getting more training data is the greatest way to minimize overfitting. Though this is not always possible in medical imaging, a model that has been trained on a larger dataset often generalizes better. Regularization with dropout or weight decay, batch normalization, data augmentation, and architectural complexity reduction are some of the other solutions. Dropout is a recently developed regularization strategy where, during training, randomly chosen activations are set to 0, making the model less sensitive to particular network weights. Weight decay, also known as L2 regularization, minimizes overfitting by punishing the model's weights, limiting their range to small values.

The risk of overfitting is reduced, gradient flow through the network is improved, greater learning rates are possible, and initialization reliance is decreased with batch normalization, a type of supplemental layer that adaptively normalizes the input values of the subsequent layer. Data augmentation is useful for reducing overfitting since it allows the model to encounter a variety of inputs during training cycles. Examples of these random transformations include flipping, translation, cropping, rotating, and random erasing of the training data. Despite these attempts, information leakage during the hyper-parameter fine-tuning and model selection process still raises the possibility of overfitting to the validation set as opposed to the training set. Therefore, it is essential for confirming the generalizability of the model to report its performance on a different (i.e., unseen) test set, and ideally on external validation datasets if applicable. As shown in Figure 2.16, monitoring the loss on the training and validation sets throughout the training iteration is a standard criterion for spotting overfitting. The model has been overfit to the training data if it performs better on the training set than the validation set. The model has been underfit to the data if it performs badly on both the training and validation sets. Although a network performs better on the training set the longer it is trained, it eventually loses its ability to generalize since it matches the training data too well [64].

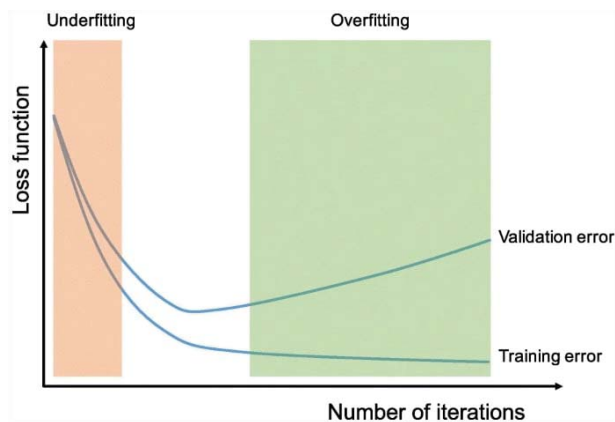


Figure 2.16: Underfitting versus overfitting (Adapted from [62]).

2.7.1.2 You Only Look Once

You Only Look Once (YOLO) is an algorithm that recognizes and finds different items in images (in real-time) [65]. The class probabilities of the discovered images are provided by the object detection process in YOLO, which is carried out as a regression problem. CNN are used by the YOLO to recognize items instantly. As the name implies, the technique only needs to detect objects once through a neural network. This indicates that a single algorithm run is used to perform prediction throughout the full image. Multiple class probabilities and bounding boxes are simultaneously predicted using the CNN. There are numerous variations of the YOLO algorithm. Tiny YOLO and YOLOv3 are a couple of the more popular ones [66]. The YOLO network is trained on the Common Objects in COntext (COCO) dataset, where the layers are divided into two 53 layers stacked on top of each other. The COCO dataset comprises 80 classes representing different types of objects [67]. The importance of YOLO resides in the following: Speed, since this algorithm can predict objects in real-time, it increases the speed of detection, high accuracy, where YOLO prediction method yields precise findings with few background errors and learning capabilities. The algorithm has outstanding learning abilities that allow it to learn object representations and utilize them when detecting the objects. YOLO algorithm employs the following three methods:

- Residual blocks
- Bounding box regression
- Intersection Over Union (IOU)

where for the residual blocks, there are several grids dividing up the image. $S \times S$ are the dimensions of each grid. For the bounding box regression, an outline that draws attention to an object in an image is called a bounding box. In order to determine the height, width, center, and class of an object, YOLO employs a single bounding box regression. The following characteristics are present in each bounding box in the image:

- Width (bw)
- Height (bh)
- Class (for example, person, car, traffic light, etc.)- This is represented by the letter c .
- Bounding box center (bx, by)

For the IOU, box overlapping is performed where IOU is used by YOLO to create an output box that properly occupies the objects. The predicted bounding boxes and their confidence scores are the responsibility of each grid cell. If the projected bounding box and the actual box match, the IOU is equal to 1. Bounding boxes that are not equivalent to the actual box are eliminated by this approach.

Regarding YOLO architecture, Figure 2.17 demonstrates the layers building up a typical YOLOv3 model. The inputs are set of images of shape $(m, 416, 416, 3)$, where m is the number of images (i.e., sample size), 416×416 is the frame size and 3 is the three channels of RGB.

These images are sent to a CNN by YOLOv3. The output's final two dimensions are flattened to provide a volume of $(19, 19, 425)$, where each cell of a 19×19 grid returns 425 numbers. For the number "425", it is 5×85 where the 5 is the number of anchor boxes per grid and the "85" is calculated from $5 + 80$, where 5 is (pc, bx, by, bh, bw) which are the class score, the x - and y - coordinates of the bounding box's top left corner and the height and the width of the bounding box, respectively. For the number "80", it is the number of classes to be detected. A list of bounding boxes and the identified classes are the output. Six integers are used to recognize each bounding box (pc, bx, by, bh, bw, c) . Each bounding box is represented by 85 values when c is expanded into an 80-dimensional vector. Finally, in order to prevent choosing overlapping boxes, the IoU and Non-Max Suppression are performed.

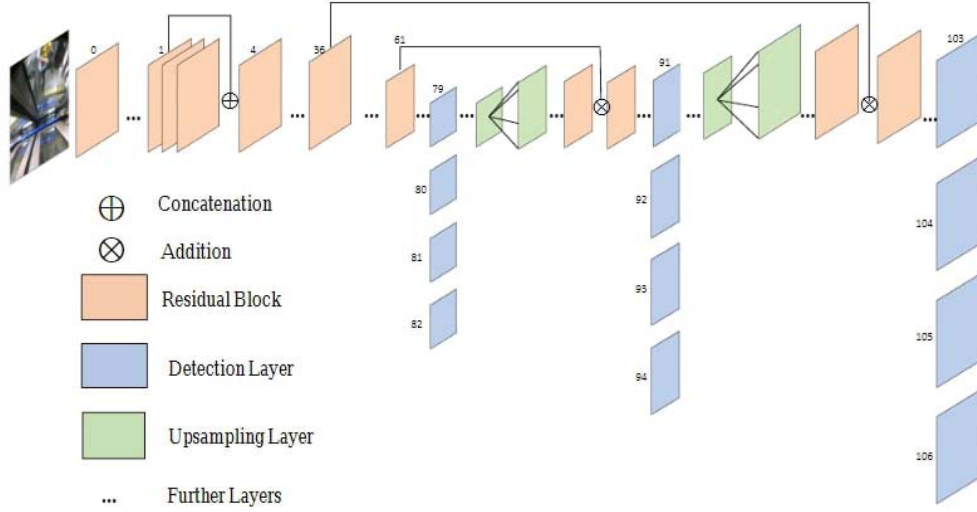


Figure 2.17: YOLOv3 network architecture (Adapted from [68]).

A 53 layer network trained on Imagenet was the basis for the Darknet variation that is used by YOLOv3. Moreover, 53 additional layers are added to it for the purpose of detection, giving YOLOv3 a 106-layer fully convolutional underlying architecture. By applying 1×1 detection kernels to feature maps of three distinct sizes located at three separate points in the network, YOLOv3 detects objects. The detection kernel has the dimensions $1 \times 1 \times (B \times (5 + C))$. Here, "5" stands for the four bounding box attributes and one object confidence, while " B " is the number of bounding boxes that a feature map cell can predict. " C " denotes the number of classes. YOLOv3 calculates the classification loss for each label using binary cross-entropy, and logistic regression is used to forecast object confidence and class predictions. Regarding the hyper-parameters used, class threshold - defines the predicted object's probability threshold. Moreover, non-max suppression threshold helps in avoiding the issue of detecting the same object repeatedly in an image. This is accomplished by selecting the boxes with the highest probability and suppressing the nearby boxes with lower probabilities (i.e., less than the predefined threshold). Finally, the image's size can be tuned according to the behaviour of the network needed.

For the convolutional layers in YOLOv3, there are 53 convolutional layers with batch normalization and leaky ReLU activation layers following each convolutional layer. The usage

2. Background

of a convolutional layer provides numerous feature maps by combining various filters with the images. The feature maps are down-sampled using a convolutional layer with stride 2, without any use of pooling. This assists in the reduction of the loss of low-level features often attributed to pooling.

2.7.1.3 Generative Adversarial Network

Generative Adversarial Networks (GANs) are a type of generative modelling that employs deep learning techniques such as CNNs. The goal of generative modeling, an unsupervised learning technique in machine learning, is to automatically identify and learn the regularities or patterns in input data so that the model may be utilized to produce or generate new examples that might have been reasonably drawn from the original dataset [69]. The GAN architecture, as shown in Figure 2.18, comprises the discriminator D that distinguishes between real and fake data, where the fake data is the data generated by the generator, and the generator G that takes an n -dimensional noise z as input, and outputs $G(z)$, which is subsequently fed as an input to the discriminator [69]. For the real data, the output of the discriminator is $D(x)$, while for the fake data, the output is $D(G(z))$. These predictions are represented in the form of probabilities P . If P is close to zero, then, this means that the input was fake, however, if it is closer to one, then, the input was real. The discriminator aims to make the probability of x as large as possible and the probability of $G(z)$ as small as possible. However, the generator aims to make the probability of the discriminator for $G(z)$ as large as possible to fool the discriminator, so that the discriminator thinks of fake data as real. This sets up an adversarial network. To rescale the output, log is taken for both $D(x)$ and $D(G(z))$, then, the mathematical expectation for both terms is also computed to get the loss function [69]. On the other hand, this loss function is complicated to solve since the discriminator has to maximize the whole term (i.e., $\mathbb{E}_{p-p(x)}\log(D(x)) + \mathbb{E}_{p-p(z)}\log(D(G(z)))$), whereas the generator tries to maximize the second term in the equation (i.e., $\mathbb{E}_{p-p(z)}\log(D(G(z)))$). So, this complexity can be reduced by subtracting $D(G(z))$ from 1 giving rise to the following loss function for GANs [69]

$$\min_G \max_D V(G, D) = \mathbb{E}_{p-p(x)}\log(D(x)) + \mathbb{E}_{p-p(z)}\log(1 - D(G(z))) \quad (2.18)$$

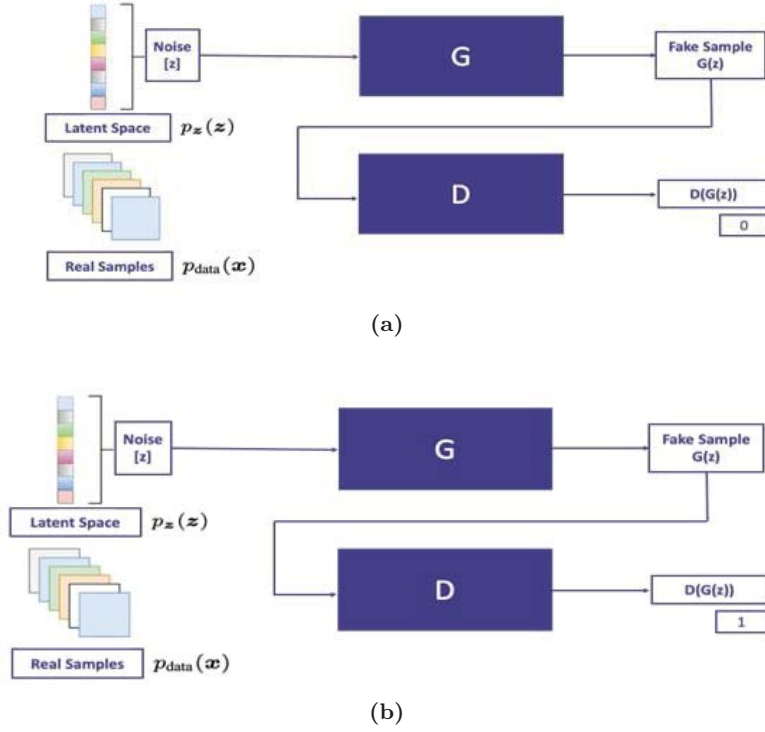


Figure 2.18: GAN architecture. (a) Discriminator perspective. (b) Generator perspective. (Adapted from [70]).

CycleGAN Architecture

The CycleGAN architecture, as shown in Figure 2.19, comprises two generators, one generator for forward mapping and the other is for inverse mapping. On both outputs, there are two discriminators connected to each of the outputs to determine whether the generated output is real or fake. For the forward mapping, the generator G tries to map an input x that belongs to a certain domain to an output y that belongs to another domain. The connected discriminator then determines whether the generated output is real or fake and penalizes the network accordingly. A second mapping generator F then takes that output and tries to reconstruct the input again (i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$). This process repeats until the network successfully fools the discriminator [71]. The error between the real input image and the reconstructed one is calculated by means of the Cycle consistency loss function. CycleGAN uses this loss function to learn the relationship between the input and the target image. Similarly, the inverse mapping is done where the generator F takes Y and tries to map it to X and a discriminator is connected to the generated output to test whether it is real or fake. Then, generator G takes the output x as an input and tries to reconstruct the input y (i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$). Since this is inverse mapping, therefore, the loss function is called backward Cycle consistency loss function unlike the normal mapping loss function that is called forward Cycle consistency loss function.

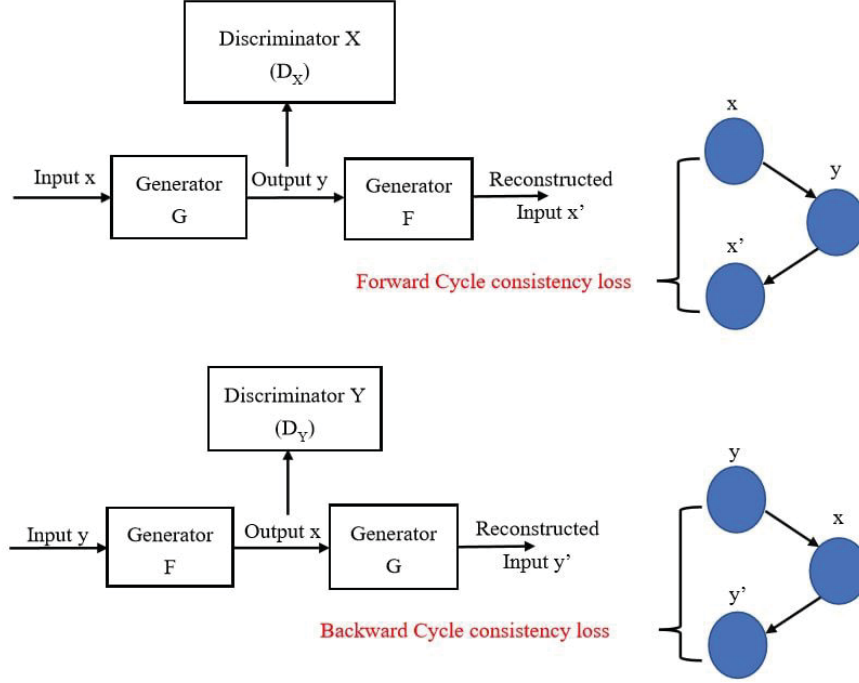


Figure 2.19: CycleGAN Architecture (Adapted from [72]).

The adversarial losses to both mappings are applied [73]. The objective function of the forward and backward mappings of CycleGAN is expressed as follows

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim P(y)}[\log[D_Y(y)]] + \mathbb{E}_{x \sim P(x)}[\log(1 - D_Y(G(x)))] \quad (2.19)$$

$$L_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x \sim P(x)}[\log[D_X(x)]] + \mathbb{E}_{y \sim P(y)}[\log(1 - D_X(F(y)))] \quad (2.20)$$

where the generators G and F try to minimize the whole term, whereas the discriminators D_X and D_Y try to maximize the whole term. For $\mathbb{E}_{y \sim P(y)}$ and $\mathbb{E}_{x \sim P(x)}$, they are the mathematical expectations for calculating the probability of y being similar to $G(x)$ and x being similar to $F(y)$. To reconstruct the input again from the output, the Cycle consistency loss function, that uses L1, is used which is expressed as

$$L_{Cycle}(G, F) = \mathbb{E}_{x \sim P(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim P(y)}[\|G(F(y)) - y\|_1] \quad (2.21)$$

So, the full objective function of the CycleGAN is expressed as

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{Cycle}(G, F) \quad (2.22)$$

where λ is controlling the relative importance of the two objectives. Finally, the main equation to solve is expressed as

$$G^*, F^* = \arg \min_{G, F} \max_{D_X, D_Y} L(G, F, D_X, D_Y) \quad (2.23)$$

The internal structures of the two generators and the two discriminators in the CycleGAN use U-Net and PatchGAN, respectively.

Pix2Pix Architecture

The Pix2Pix GAN uses conditional GAN for image-to-image translation. The difference between conditional GANs and GANs is that in GANs, there is no control over the modes of the generated data in GANs, whereas the conditional GAN generates images conditioned on a class label [74]. A block diagram of the Pix2Pix approach that is used in image-to-image translation, is shown in Figure 2.20 [75]. A photo, that is composed of two images concatenated together, is entered as an input to the generator to predict an image similar to the target domain. The left image in the input image is the clip art image from the target domain and the right image in the input image is the image from the real-high resolution photos domain. The generator loss is then calculated from the prediction and optimized to fool the discriminator. The loss is measuring the similarity of the predicted image in comparison to the original target image. However, the discriminator takes as input the original target image along with the predicted image by the generator to figure out if the generated image is indistinguishable from the original target image or not. The discriminator loss is then calculated and optimized to train the discriminator to figure out fake outputs generated by the generator [76].

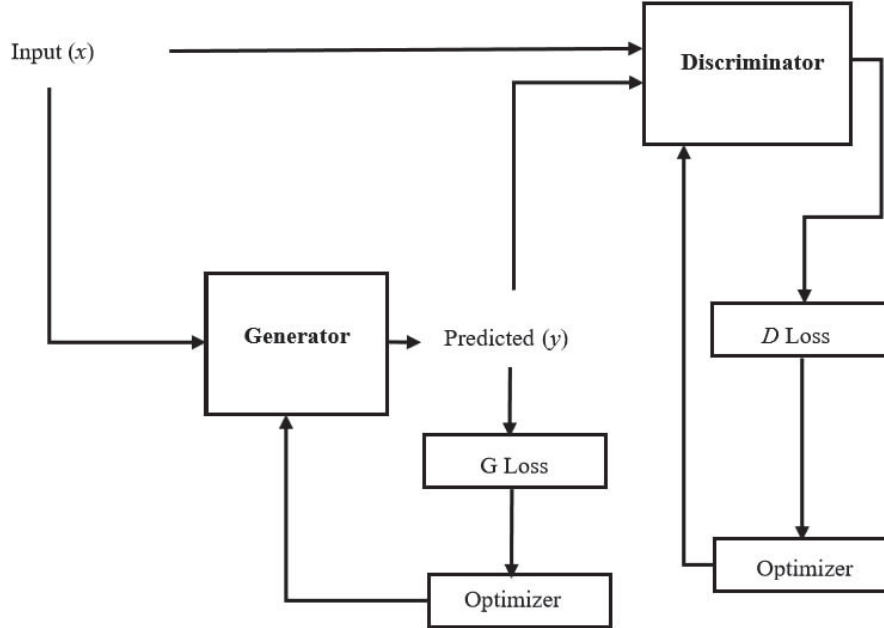


Figure 2.20: Pix2Pix GAN architecture (Adapted from [75]).

To illustrate how Pix2Pix is used, the generator and discriminator roles remain the same as that of any GAN model. However, the difference is that the generator uses a U-Net-based architecture which is a rapid and precise segmentation of images, whereas the discriminator uses patchGAN classifier that tries to classify each $N \times N$ patch in an image to be real/fake as shown in Figure 2.21a and Figure 2.21b, respectively. A loss function is calculated between the output generated from the generator and the original image until the generator is able

to produce indistinguishable clip art images from the original clip art images to deceive the discriminator. Back-propagation is applied to update the weights using Gradient Descent Optimizer [77].

Pix2Pix uses conditional GAN for image-to-image translation. Accordingly, it has the same objective function of the conditional GAN that can be expressed as [75]

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log(D(x, y))] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \quad (2.24)$$

where $L_{cGAN}(G, D)$ is the adversarial loss between the generator and the discriminator. The variables x , y and z in the equation are the input image, target image (i.e., ground truth image) and random noise vector, respectively [75]. $D(\cdot)$ is the probability that the generated/target image is real when compared to the photo x , and $\mathbb{E}_{x,y}$ and $\mathbb{E}_{x,z}$ are the mathematical expectations of $\log(D(x, y))$ and $\log(1 - D(x, G(x, z)))$, respectively. In this loss function, the discriminator attempts to tune its parameters to improve its ability in discriminating the real input (Class 1) from the generated fake image (Class 0) (i.e., increase $D(x, y)$ to indicate that y is the real target image and decrease $(D(x, G(x, z)))$ to indicate that the generated image by the generator is far from the target image), thus, maximize $L_{cGAN}(G, D)$. On the other hand, the generator attempts to tune its parameters to deceive the discriminator to classify a generated fake image as belonging to Class 1 (i.e., increase $D(x, G(x, z))$). Given the representation of $L_{cGAN}(G, D)$ in Equation 2.24 which includes the term $(1 - D(x, G(x, z)))$, the generator thus attempts to minimize the tuning of the generator would have no impact on the first term that includes $D(x, y)$. This leads to identifying the optimal generator G^* as

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) \quad (2.25)$$

This equation will successfully fool the discriminator; however, the main objective now is to generate an image that is close to the ground truth output. Therefore, regularization needs to be added to penalize the network every time it produces an undesired output [78]. In this case, the discriminator's job remained unchanged. However, the generator is tasked to not only fool the discriminator, but also to be close to the ground truth output. This is implemented using L1 distance as

$$L_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \quad (2.26)$$

So, the final objective is given by

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (2.27)$$

where λ regulates the relative significance of the two objectives and the generator learns to ignore the noise that was generated before introducing L1 [79].

The generator architecture, as shown in Figure 2.21a, employs U-Net, which is an encoder-decoder model with skip connections between reflected layers in the encoder and the decoder stacks to allow low-level information to shortcut across the network [80]. In addition, Figure 2.21b shows the PatchGAN classifier used in the discriminator, where instead of indicating

2. Background

that a generated image is real or fake, the discriminator determines whether an $N \times N$ patch of the generated image is real or fake [81].

The internal structures of the generator and the discriminator are as follows: The generator is composed of 8 convolutional layers for the down-sampling portion of the U-Net, where each layer, in addition to convolutional filters, has batch normalization and Leaky ReLU activations. The number of filters used in the down-sampling part in each layer is 64, 128, 256, 512, 512, 512, 512 and 512, respectively. The filter size is 2×2 for each layer. The up-sampling portion comprises 7 convolutional layers, where each layer also has batch normalization and ReLU activation. The number of filters used in the up-sampling part in each layer is 512, 512, 512, 512, 256, 128 and 64, respectively.

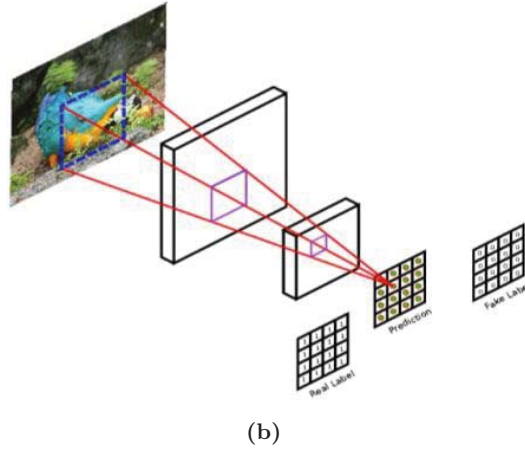
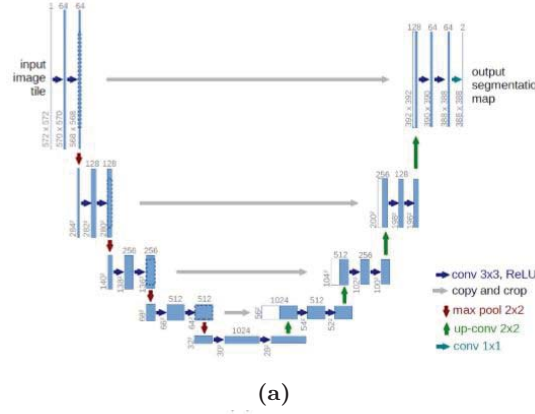


Figure 2.21: Generator and discriminator architectures. (a) U-Net Architecture used in the generator (Adapted from [80]). (b) PatchGAN classifier used in the discriminator (Adapted from [81]).

The filter size is 2×2 for each layer. In addition, skip-connection is performed by concatenating layers in down-sampling with up-sampling portion. Furthermore, the binary cross-entropy loss function is utilized for the generator and discriminator. On the other hand, the discriminator comprises 3 convolutional layers, each followed by batch normalization and Leaky ReLU. The number of filters used in the discriminator in each layer is 64, 128 and 256, respectively, with a filter size of 2×2 for each layer. Zero-padding is then applied to keep the specified shape of the output. Another convolutional layer and Leaky ReLU layer are added along with a zero-padding layer. Finally, Adam optimizer is utilized for the generator and discriminator.

In addition to the conceptual differences between CycleGAN and Pix2Pix, the difference in architecture between the CycleGAN and Pix2Pix is that CycleGAN uses instance normalization in each of the two generators and two discriminators, instead of batch normalization that is used by Pix2Pix [71].

2.7.2 Related Work

In order to overcome some of the major limitations of blindness, computer vision plays a crucial part in prosthetic vision [82]. Sanchez-Garcia et al. propose a novel method for creating a schematic representation of interior environments using phosphene images. The structural characteristics in a scene are extracted using computer vision and deep learning algorithms, and various indoor environments created for prosthetic vision are recognized. The proposed technique relies on the extraction of structural informative edges, which can support a variety of computer vision tasks like object recognition and scene understanding by clearly expressing the scene’s structure. A method for object detection that makes use of a precise machine learning model, is employed that can locate and recognize several objects in a single image. A phosphenes pattern was utilized to represent the extracted information. A fully convolutional network was utilized to perform pixel-wise labelling that learns to anticipate “structural informative edges” for various indoor scene structures. Eleven participants volunteered to test the method’s efficacy using real data from indoor settings. The findings demonstrate that, even at low resolutions, the newly proposed technique gave an accuracy of 58% in the recognition of indoor environments.

A novel method for creating a schematic illustration of interior settings for simulated phosphene images, was outlined [83]. The suggested approach by Sanchez-Garcia et al. integrates a number of convolutional neural networks to convey and extract relevant information about the scene, such as structurally informative edges of the surroundings and silhouettes of segmented objects. Normal sighted test subjects were used in experiments using a Simulated Prosthetic Vision system. In comparison to previous image processing techniques, the results of employing the suggested strategy for interior scenes demonstrate 95% confidence intervals and 97 % for the recall and the precision for object recognition and room identification tasks.

Face information is crucial for identifying individuals, but it is difficult to discern at low resolution [84]. The goal of a psychological physics experiment on person recognition in daily life was to find the most effective way to employ the few available stimulating electrodes while still providing valuable visual information. By converting complex face information into simple Chinese character information, Zhao et al., employ the real-time image-processing technique based on FaceNet to optimize the person information. It was observed that the target individuals’ last names were hidden under different Chinese characters that covered every processed target face. According to the psychological findings, the FaceNet-based image-processing method increased recognition accuracy to 100 % when the resolution increased to 64×64 . The suggested method, which involves converting complicated facial information into simple Chinese character information, would enable participants to use their own prior knowledge to identify the required person more quickly and reliably.

Despite the tremendous improvements in deep learning, it is still impossible for researchers to come up with a general, automated pre-processing method that can be customized for certain

applications or user needs [85]. Van Steveninck et al. propose a novel deep learning approach that directly addresses this problem by end-to-end optimizing the phosphene synthesis process. The suggested model features a highly configurable simulation module for prosthetic vision and is based on a deep auto-encoder architecture. It is demonstrated that such a technique is capable of automatically finding a task-specific stimulation protocol in computational validation trials. The outcomes of these proof-of-concept tests demonstrate the promise of end-to-end prosthetic vision optimization. An accuracy of 69.7% for the boundary pixels in the supervised semantic boundary condition was achieved. The proposed approach is fairly flexible, and it could be expanded to automatically dynamically optimize prosthetic vision for everyday tasks under any given limits while taking into account the specific needs of the end user.

3

Image Enhancement and Phosphene Simulation

3.1 Image Enhancement

To enhance the perceived images by visual prostheses users, image enhancement techniques are proposed in this thesis to help in object recognition enhancement. The proposed enhancement techniques were applied in phosphene simulation experiments to mimic the environment that a typical visual prosthetic user encounters. The proposed enhancement techniques are employed as image pre-processing in the dropout handling proposed approach in Chapter 4 and in the scene simplification proposed approach in Chapter 5. The input image undergoes the following image enhancement steps:

- Take the image as an input where the primary color model is the “RGB” color space.
- Resize the image to be of size 32×32 which has been previously considered the threshold of recognition tasks and scene perception [25].
- Convert the resized image to grayscale since visual prostheses’ users perceive grayscale images.
- Since it was reported that high-contrast affects the perception of images [18], apply CLAHE to improve the contrast of the image.
- Apply Wiener filter of size 3×3 to remove the noise [86].
- Apply Otsu thresholding to perform automatic image thresholding where a threshold value is calculated, based on the variance, to determine which pixels will be marked as foreground and which pixels will be marked as background [87].
- Apply a Median filter of size 3×3 to remove noise and sharpen the edges more to maintain the image’s details [56].

- Invert the colours of the foreground and the background in case the foreground is black.
- [Optional] Apply Dilation to increase the ratio of the foreground with respect to the background.
- Apply connected components labelling using 8 connectivity.
- Discard any region(noise) that has an area less than a specific area values using false-positive reduction techniques.
- Apply phosphene simulation for the image.

The foreground should be always brighter than the background, so that the object of interest will be easily recognized in the phosphene simulation. Equation (3.1) shows the calculation for the summation of pixels in the corners of the image (i.e., pixels in the first and last rows and in the first and last columns). Equation (3.2) shows the calculation of the intensity value to determine whether the pixel is a foreground or a background pixel. In case the foreground is black (i.e., $\text{round}(SoC) \neq 0$), the foreground and background colors will be inverted so that the foreground will be brighter than that of the background when used in phosphene simulation.

$$SoC = SRF + SRL + SCF + SCL \quad (3.1)$$

where SoC is the summation of the pixels at the corners. However, SRF , SRL , SCF and SCL are the summation of the first row pixels, the last row pixels, the first column pixels and the last column pixels, respectively.










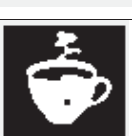
$$r = \frac{SoC}{2^n} \quad (3.2)$$

where r is the normalization of SoC where SoC is taken from Equation (3.1) and n is the number of bits used to represent the pixels' intensities.

To illustrate how the noise in any image was discarded, we followed the box plot functionality where the first quartile is named Q_1 and the third quartile is named Q_3 [88]. By means of connected components labelling using 8-connectivity, the objects in an image are extracted. Thus, the area of any of the extracted objects should be within the first quartile (i.e., 150) and the third quartile (i.e., 256) so that the corresponding object will not be identified as noise and therefore, discarded [89].

Two enhancement techniques were proposed in this thesis, one with dilation and the other one without dilation, to examine the more efficient enhancement technique. In Table 3.1, the image enhancement is examined both with performing dilation before the removal of unwanted regions and sampling the image to be 32×32 and also without dilation.

Table 3.1: Proposed image processing techniques with and without dilation.

Image Enhancement Technique	Output Image	Comments
Input Image		
Grayscale Conversion		
CLAHE Applied		
Wiener Filter Applied		
Otsu Thresholding		
Median Filter		
Colors Inverted		No conversion is needed since the foreground color is already white
Dilation Applied		
Removal of unwanted regions		with Dilation applied
Removal of unwanted regions		without Dilation Applied

3.2 Phosphene Simulation

Since adding dilation thickens the objects in the image, we proceed without using dilation to preserve the difference between foreground and background in the image. The main features proposed in the phosphene simulation for the ideal phosphenes type are the phosphene shape whether it is square, circle or hexagon where the shape codes are 1, 2 and 3 each corresponding to square, circle and hexagon, respectively; the distance between each two successive phosphenes; the grid that the phosphenes simulation will follow whether the squared grid or the hexagonal grid; the resolution of the grid.

Table 3.2 shows the ideal phosphene simulation with square-shaped, circular-shaped and hexagonal-shaped phosphenes. Moreover, the table shows the distance between two successive phosphenes (i.e., 0 or 0.5), grid utilized (i.e., squared or hexagonal grids) and grid resolution 32×32 . It can be observed that the more the distance between each two successive phosphenes, the less the size of the phosphene. Moreover, the utilization of either hexagonal grid or squared grid, gave the same ability in recognizing the cup with a smoke coming out of it, which justifies that the use of regular grid, helps in object recognition. Finally, it can be observed that using circular-shaped phosphenes, gives a better demonstration of the phosphene definition which is a spot of light.

Table 3.2: Ideal phosphene simulation with no dilation.

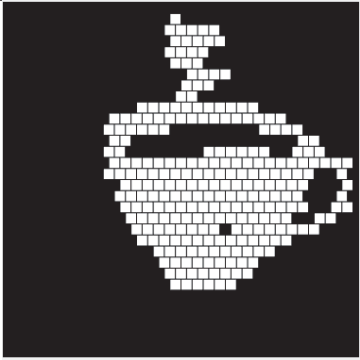
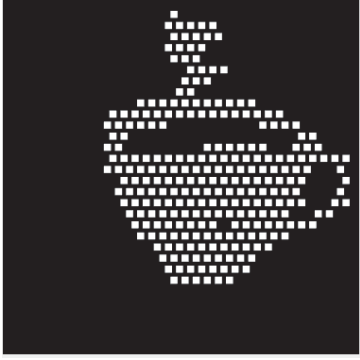
Phosphene shape code	Desired in between Distance	Grid used	Grid Resolution	Phosphene simulation Image
1	0	Hexagonal Grid	32 x 32	
1	0.5	Hexagonal Grid	32 x 32	
Continued on next page				

Table 3.2 – continued from previous page

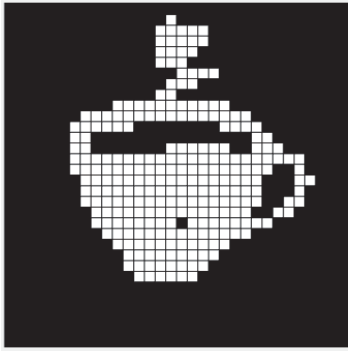
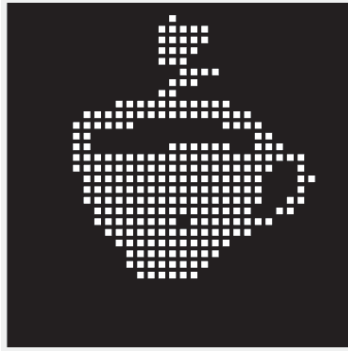
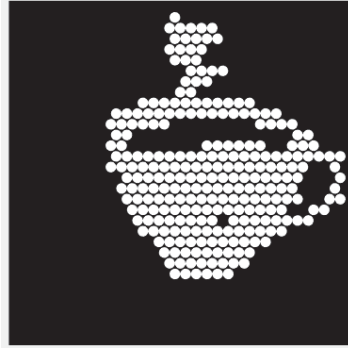
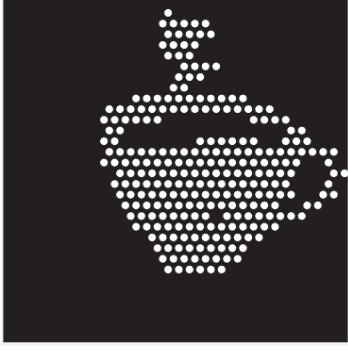
Phosphene shape code	Desired in between Distance	Grid used	Grid Resolution	Phosphene simulation Image
1	0	Squared Grid	32 x 32	
1	0.5	Squared Grid	32 x 32	
2	0	Hexagonal Grid	32 x 32	
2	0.5	Hexagonal Grid	32 x 32	
Continued on next page				

Table 3.2 – continued from previous page

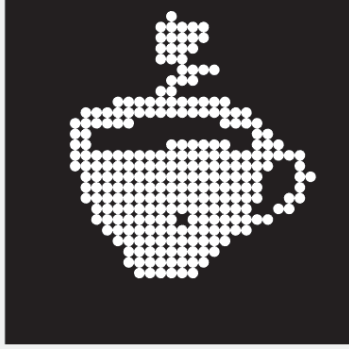
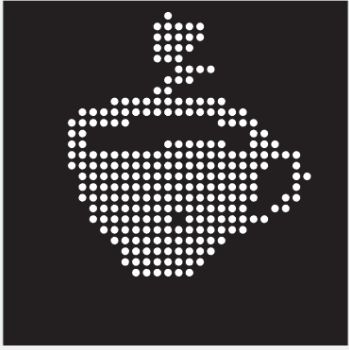
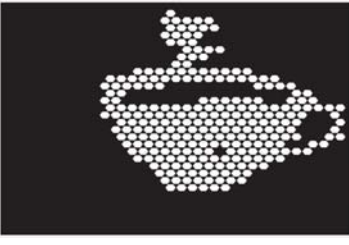



Phosphene shape code	Desired in between Distance	Grid used	Grid Resolution	Phosphene simulation Image
2	0	Squared Grid	32 x 32	
2	0.5	Squared Grid	32 x 32	
3	0	Hexagonal Grid	32 x 32	
3	0.5	Hexagonal Grid	32 x 32	
3	0	Squared Grid	32 x 32	
Continued on next page				

Table 3.2 – continued from previous page

Phosphene shape code	Desired in between Distance	Grid used	Grid Resolution	Phosphene simulation Image
3	0.5	Squared Grid	32 x 32	

In this thesis, three phosphene simulation techniques were tried which are ideal phosphene simulation, scoreboard phosphene simulation and axon map phosphene simulation. The size of phosphenes used was adjusted relative to the distance between each two successive phosphenes. Moreover, the distance between each two successive phosphenes was set to 0.5 to match the placement of electrodes in the implant [90]. The phosphene shape used was circular shape to match the common form for phosphenes simulating the actual look of the phosphenes without any change in the stimulation amplitude and with ideal current set [91]. The grid used was squared grid since it simulates the common grid used in visual prostheses [28]. Since visual prostheses users perceive grayscale images, we implemented grayscale phosphene simulation. The number of gray levels that was reported in multiple studies by visual prostheses implanted patients ranges from 4 to 12 levels [28, 92–94]. In addition, other studies reported that reaching 8 to 16 gray levels could be achieved [95]. Therefore, we used 8 gray levels in our simulations to match the mid-range of the values reported in the literature [96]. Finally, the resolution used was 32×32 which is the threshold for scene recognition [25, 97].

All the experimental procedures presented in this thesis, were approved by the Faculty of Media Engineering and Technology, German University in Cairo from the ethical standpoint and is in accordance with the ethical standards of the Declaration of Helsinki. All participants involved in the experiments signed an informed consent form. This was performed for both the computer screen experiments and the mixed reality (MR) experiments.

All the computer screen experiments presented in this thesis were performed on corrected vision/normally-sighted subjects seated on a chair facing a 15-inch computer screen at 1m distance. This results in a 20° simulated field of view, which is the limit for legal blindness [25], as shown in Figure 3.1.

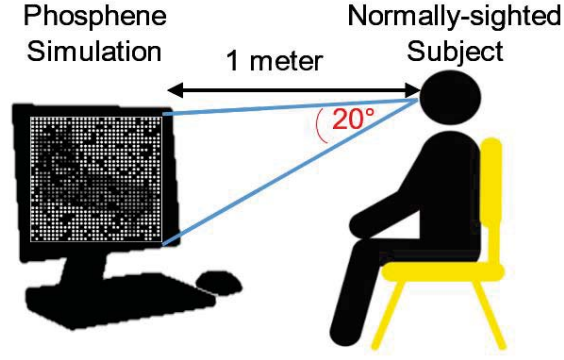


Figure 3.1: Experimental setup.

A demonstration that shows how a real visual prosthesis system, utilizing the proposed approaches discussed in this thesis, is shown in Figure 3.2. Figure 3.2 shows that the proposed algorithms in this thesis are to be processed inside the VPU in a typical visual prosthesis system that includes the proposed dropout handling technique and the proposed scene simplification technique using either YOLO or GAN. Then, the processed information, after being transformed to electrical signals, will be transferred wirelessly to a receiver in the implant enabling the wearer to perceive the external environment with the enhancement applied.

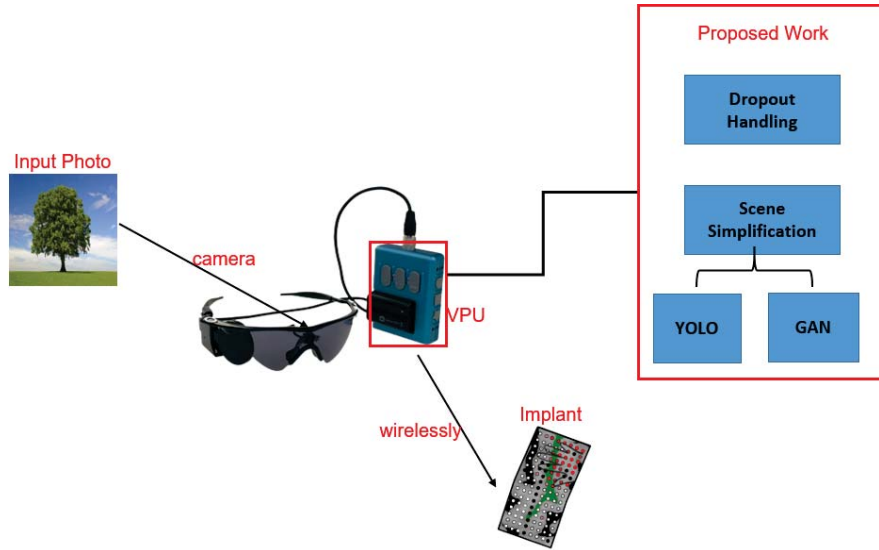


Figure 3.2: An overview of a real visual prosthesis system utilizing the proposed techniques.

4

Electrode Dropout Handling

4.1 Introduction

A major challenge that reduces the quality of the perceived image in prosthetic vision is stimulation electrode dropout. This occurs due to the malfunction of electrodes post-implantation or being implanted in dead tissue, which causes permanent black spots in the location of the corresponding phosphene(s) [98]. This negatively affects the recognition and identification of objects since some of the key-points in the object may be aligned with the receptive field of the neurons stimulated by the dropped-out electrodes. The most frequently observed percentages of phosphenes' dropouts have been shown to range from 10% to 30%, as shown in Figure 2.7 [99, 100].

We propose an image processing approach to compensate for the effects of electrode dropout on the perceived image. In this approach, convolution between a bounding box of the object of interest and the phosphene grid is used to identify and translate the object of interest to a location within the image that fits the object and has minimum dropouts. This is performed after showing the actual location of the object of interest to help in accurately identifying its location. We examined the performance of the proposed approach on different groups of normally sighted subjects using simulated prosthetic vision. Despite the differences between simulated and actual prosthetic vision, where actual prosthetic vision tends to be more complex, we adopted a phosphene simulation strategy that mimics perceived images reported by visual prostheses users [28, 91]. The results indicate a significant enhancement in recognizing the objects of interest using the proposed dropout handling approach. This could eliminate the need for performing additional interventions to replace the malfunctioning electrode.

4.2 Methods

4.2.1 Image Pre-processing

To allow better and faster identification of the object's identity in any prosthetic vision scene in the presence of dropouts, we propose an approach that optimally translate the object

in the scene to a new location that has minimum number of dropouts. In this approach, the input image undergoes preprocessing prior to applying the dropout handling mechanism. The pre-processing performed follows the same proposed enhancement technique discussed in Chapter 3.

4.2.2 Phosphene Simulation

The output of the pre-processing stages is used to generate a simulated prosthetic vision image. We used a square grid for phosphene representation [91]. A circular phosphene shape is also used in this study; consistent with multiple simulated prosthetic vision studies [28]. The distance between each two consecutive phosphenes was set to zero. Dropouts were then simulated with rates of 10%, 20% or 30% of the total number of pixels [100]. The location of the dropped out phosphenes was determined randomly following a uniform distribution. In all simulations, a dropped out phosphene was set to a black color.

4.2.3 Dropout Handling Approach

The main aim is to minimize the impact of dropouts on the perception. This is done by translating the object in a certain image to a certain location in the same image where the number of dropped out phosphenes is minimum. In this approach, we construct a matrix D of size 32×32 (i.e., the same enhanced input image size) in which dropouts are represented by 0, while other pixels are represented by 1. A bounding box B is then identified around the object of interest by computing the connected-components labeling using 8-connectivity to get the connected pixels that refer to the same object and retrieve the minimum row, the minimum column, the maximum row and the maximum column. All pixels within B are set to 1. A convolution process is then performed between B and D , and the optimal location for the center of the object of interest C_{New} is identified by the position that has the highest value in the convolution output defined as

$$C_{New} = \max_{(x,y)} \{(B * D)[x, y]\} \quad (4.1)$$

Finally, we translate the object within B in the presented image to C_{New} . If multiple locations have the same maximum value in the convolution output, C_{New} is set as the location that is closest to the center of B based on the Euclidean distance.

4.2.4 Experimental Design and Procedure

To examine the efficacy of the proposed approach, 12 subjects participated in computer-screen experiments (5 males and 7 females) of age 22 to 60 years. Prior to the presentation of the images, each subject was given a demonstration that included 3 different test images that are different from the images used in the following experiments to avoid any learning effects. Two versions of the same image were used in the demonstration: The first version is displayed in terms of phosphenes to introduce the subjects to prosthetic vision, and the second version demonstrated the effect of phosphene dropout (i.e., black spots).

The test subjects were divided into 4 groups with 3 subjects each. Each group participated in a different experiment in which each subject was presented with 24 different test images: 8

images with 10% dropout, 8 images with 20% dropout and 8 images with 30% dropout. The 24 test images set was fixed across all subjects. Each test image was displayed in a trial of duration 10 sec and then the subjects were given the chance to tell the identity of the object displayed in the image. The images represented objects from different categories including car, utensils, flower, window, bed, chair, numbers, bird, teapot, stairs, shelf, and truck.

The first group of subjects was presented with the test images after performing the pre-processing and prosthetic vision simulation for the whole 10 sec. No dropout handling was performed. This group represents what current visual prostheses users would perceive. The second group was presented with the same image presented to the first group for 5 sec only. The object is then translated to a random position within the image and displayed for another 5 sec. This test was performed to determine whether any enhancement in performance was in fact due to the proposed dropout handling approach or as a result of displaying two different versions of the object to the subject. Dropout handling was examined in the two other groups. The third group was presented with the output of the proposed dropout handling approach for the entire 10 sec. The last group was presented with the test images after pre-processing and prosthetic vision simulation without dropout handling for 5 sec, and then with the output of the dropout handling approach for another 5 sec. This last group represents, contrary to the third group, a practical implementation of the proposed approach as it still provides the subject with the actual location of the object of interest during the first 5 sec, and then a better presentation of the object for another 5 sec after dropout handling. This could help in providing the subject with an accurate localization of the object while minimizing the effects of dropout.

4.2.5 Evaluation Metrics

Three metrics were used to assess the performance of each group. First, we measured the object recognition accuracy as

$$Accuracy(\%) = \frac{N}{T} \times 100 \quad (4.2)$$

where N is the number of correctly identified objects by the subject and T is the total number of presented test images. Second, the time taken by the subject to correctly recognize the presented object was recorded, with a maximum of 10 sec. Finally, the subjects reported their confidence level in the recognition on a scale of 1 to 5.

4.3 Results

4.3.1 Dropout Simulation and Handling

We first demonstrate the output of each stage of the proposed approach. Figure 4.1 illustrates an overview of the process implemented to show the effect of the dropout handling approach on the object clarification and, thus, object recognition. It shows a sample test image representing the number “eight” for different dropout rates. The figure demonstrates, especially in the case of 30% dropout rate, how translating the object of interest to a location with minimum number of dropouts enhances the presentation.

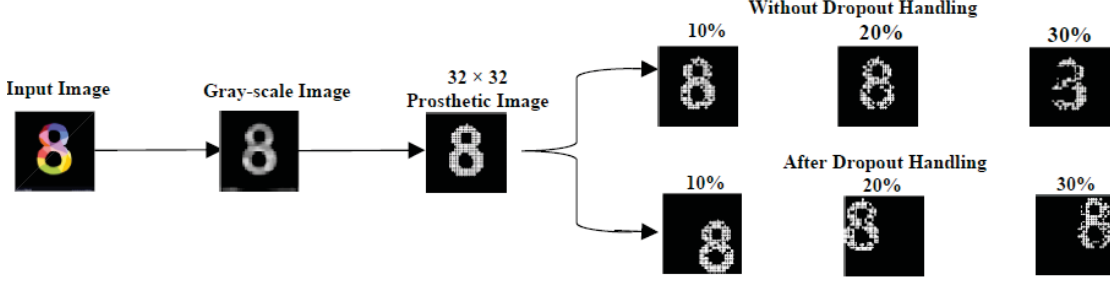


Figure 4.1: Dropout simulation and handling for different dropout rates of 10%, 20% and 30%.

4.3.2 Performance Evaluation

We examined the utility of the proposed approach by presenting different sets of images to four groups of subjects. For each group, images with 10%, 20% and 30% dropout rates were presented. Figure 4.2 illustrates the performance of the test subjects for a dropout rate of 10%. Figure 4.2a demonstrates significantly higher recognition accuracy when the proposed dropout handling approach was used compared to not applying dropout handling and compared to presenting the subjects with the images without dropout handling for 5 sec in addition to a randomly translated version of the image for another 5 sec (No dropout handling: $66.67 \pm 29.56\%$, Random placement: $77.1 \pm 19.79\%$, Dropout handling: $97.91 \pm 5.9\%$, $P < 0.05$, $n = 24$, two-sample t-test). Moreover, the figures demonstrate that dropout handling results in significantly shorter recognition time (Figure 4.2b; No dropout handling: 8.68 ± 1.55 sec, Random placement: 7.88 ± 2.38 sec, Dropout handling: 3.33 ± 1.28 sec, $P < 1e - 07$, $n = 24$, two-sample t-test) and higher decision confidence (Figure 4.2c; No dropout handling: 3.48 ± 0.79 , Random placement: 3.75 ± 0.74 , Dropout handling: 4.91 ± 0.24 , $P < 1e - 04$, $n = 24$, two-sample t-test).

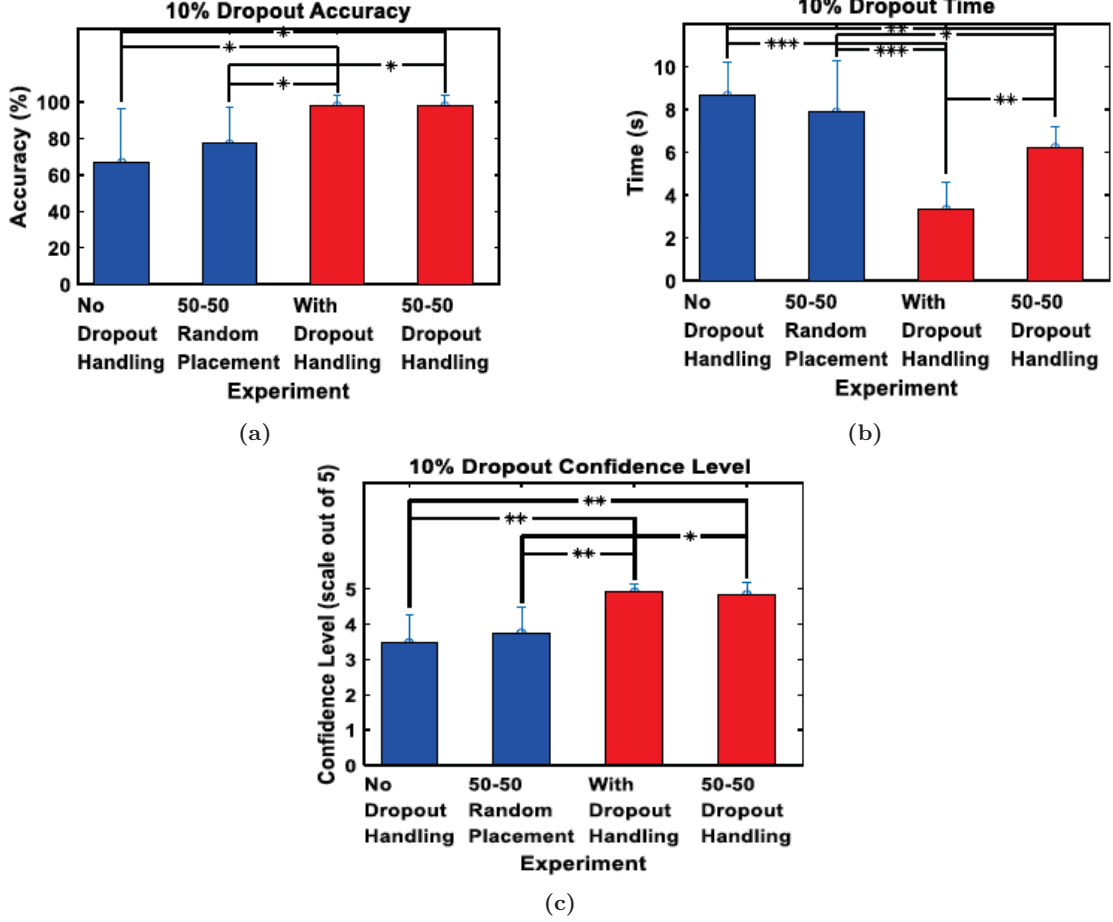


Figure 4.2: Results for 10% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level (*mean \pm std*). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

One disadvantage of applying the proposed approach is that it modifies the actual location of the displayed object. This might be confusing to a visual prosthesis user as it does not accurately represent the visual field of the subject. Therefore, to provide the subjects with both the actual location of the object in addition to a better representation with least dropouts, we modified the image presentation for the last group of subjects. In this presentation, the image is presented without dropout handling for 5 sec followed by presenting the image after dropout handling for another 5 sec. A significantly better performance can still be observed using this approach compared to no dropout handling and the random placement approaches (Recognition accuracy: $97.9 \pm 5.9\%$, Time: 6.24 ± 0.94 sec, Confidence level: 4.84 ± 0.35 , $P < 0.05$, $n = 24$, two-sample t-test). In addition, no significant difference can be observed between this modified presentation and applying the dropout handling for the entire 10 sec in terms of both the recognition accuracy and the confidence level. However, in terms of the time taken to decide the identity of the object, higher time was needed by the subjects. This is expected given that for half of the time (i.e., 5 sec), the dropout handling approach was not applied, which contributes to the time needed to recognize the object.

Examining the performance of the four groups of subjects with dropout rates of 20%

and 30% as demonstrated in Figure 4.3 and Figure 4.4, respectively, revealed a consistent enhancement when using the proposed dropout handling approach. First, significantly higher recognition accuracy can be observed using the dropout handling approach applied for the entire 10 sec compared to not applying dropout handling and the random placement approaches in the case of 20% dropout rate (No dropout handling: $62.5 \pm 44.32\%$, Random placement: $58.34 \pm 41.79\%$, Dropout handling: $95.83 \pm 7.73\%$, $P < 0.05$, $n = 24$, two-sample t-test). For the 30% dropout rate, higher, yet not statistically significant, recognition accuracy can be observed using dropout handling (No dropout handling: $87.5 \pm 35.35\%$, Random placement: $83.34 \pm 35.63\%$, Dropout handling: 100%). The lack of significance in this case can be explained given that for 30% dropout rate, relatively simpler test images were used to compensate for the high dropout rate used. Second, enhancement in the performance can be observed in both the time taken to decide the identity of the presented object and the decision confidence for a dropout rate of 20% (Time: No dropout handling: 8.93 ± 1.51 sec, Random placement: 8.41 ± 2.22 sec, Dropout handling: 5.08 ± 1.69 sec, $P < 1e - 07$, $n = 24$, two-sample t-test; Confidence level: No dropout handling: 3.28 ± 1.25 , Random placement: 3.2 ± 1.6 , Dropout handling: 4.63 ± 0.7 , $P < 1e - 04$, $n = 24$, two-sample t-test) and a dropout rate of 30% (Time: No dropout handling: 8.33 ± 1.28 sec, Random placement: 7.08 ± 1.53 sec, Dropout handling: 3.86 ± 0.93 sec, $P < 1e - 07$, $n = 24$, two-sample t-test; Confidence level: No dropout handling: 3.95 ± 1.32 , Random placement: 4.16 ± 1.24 , Dropout handling: 4.96 ± 0.11 , $P < 1e - 04$, $n = 24$, two-sample t-test). Finally, applying the dropout handling approach for 5 sec only out of the entire 10 sec showed no significant difference in both the recognition accuracy and the confidence level compared to using the dropout handling approach for the entire 10 sec for both 20% and 30% dropout rates. However, similar to the 10% dropout rate, significant increase in the time taken to decide the identity of the object can be observed.

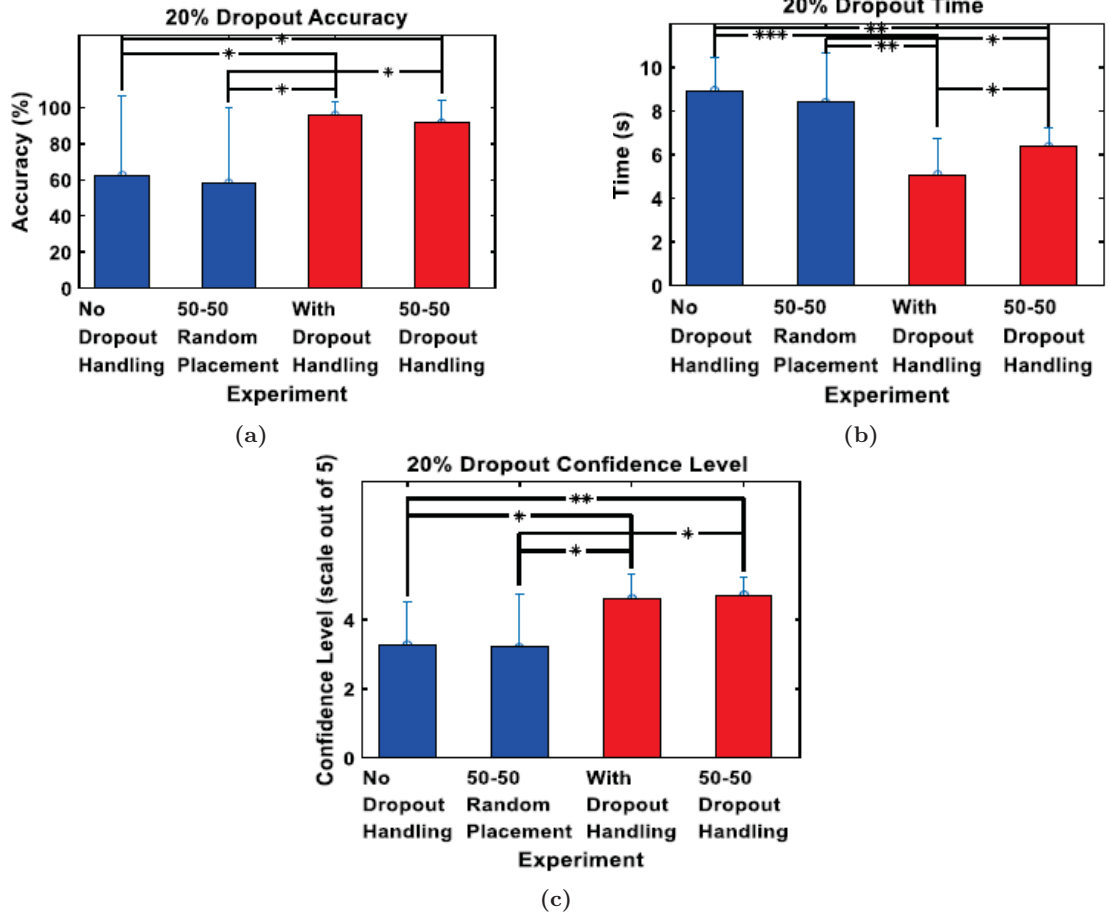


Figure 4.3: Results for 20% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level (*mean* \pm *std*). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.

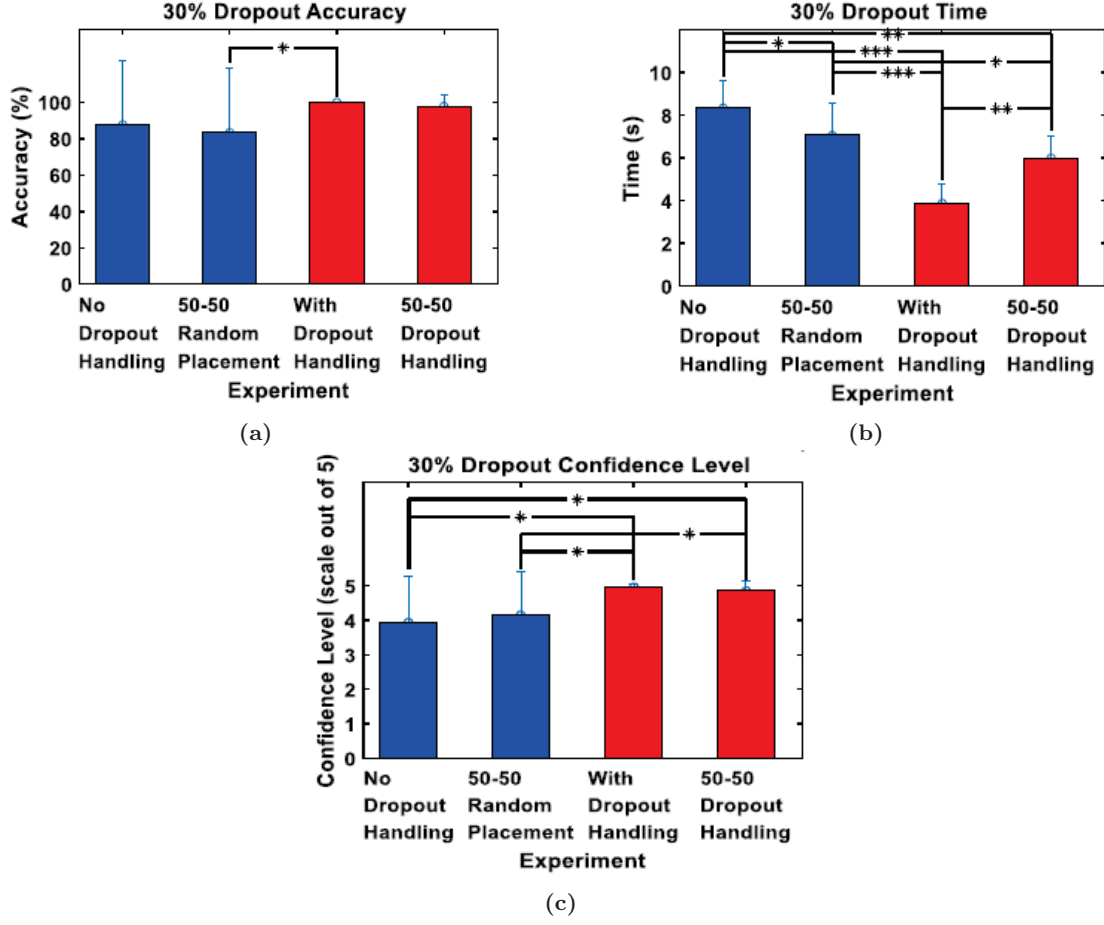


Figure 4.4: Results for 30% dropout rate for three metrics (a) recognition accuracy, (b) time to decision, and (c) confidence level (*mean \pm std*). Blue bars represent not using dropout handling, while red bars represent two different versions of dropout handling. * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

4.4 Conclusion

Visual prostheses have recently demonstrated success in restoring vision to the blind. One of the challenges that affect the quality of the perceived prosthetic image is electrode dropout. We proposed an approach that could help in better recognition of the identity of an object by means of convolution and translation. The object's phosphenes are optimally translated to a location with minimum number of electrodes dropouts. Experiments using simulated prosthetic vision revealed a significant enhancement in the ability of the test subjects to recognize the presented objects using the proposed approach compared to presenting the test image without dropout handling to the subjects as well as randomly translating the object within the image. Additionally, for the practical utilization of the approach, presenting the output of the proposed approach to the test subjects subsequent to presenting the image without dropout handling showed similar enhancement. These results indicate the efficacy of the proposed approach in compensating for the effects of electrode dropout in visual prostheses.

5

Scene Simplification

Object recognition in visual prosthesis is considered a challenge that is reported by visual prosthetic users [101]. This affects their independence level as they might need to rely on other means, such as audio for example, to be able to determine the object they are looking at. The main reason for the difficulty of object recognition in this case is the low resolution of the perceived image that reduces the details in the image. To enhance object recognition for visual prostheses' users, we propose the utilization of two deep learning-based approaches to simplify the objects in a scene; one using YOLO and the other using GAN. Clip art is a simple artwork abstraction of an object. So, the utilization of clip art in a low resolution environment, as in visual prostheses, will preserve the identity of the object while keeping its representation simple.

5.1 YOLO-Based Scene Simplification

5.1.1 Introduction

By means of a deep learning approach which is based on YOLO, recognition of the objects in a certain image can be performed [65]. The proposed approach provides an automated system that takes as input the photo and then, performs object recognition using YOLO. Subsequently, the label (i.e., the object's identity) of the detected object is identified and the corresponding clip art representation is used as input to the visual prosthesis. We evaluated the performance of the system by performing three experiments conducted on normally sighted subjects of different age groups using simulated prosthetic vision. To mimic the perceived images by the visual prostheses users, we used a phosphene simulation strategy to represent the visual prosthetic environment. The results from the experiments show that relying on the clip art representation of images as determined through YOLO-recognition allowed the subjects to understand the scenes easier and with more confidence.

5.1.2 Methods

To illustrate the steps performed in the proposed approach, Figure 5.1 shows a block diagram that inputs the photo of the scene to YOLO to get the labels of the objects in the photo. Next, each of the labels is automatically input to the search bar in Google Images to get the corresponding clip art of each of the detected objects in the photo. Pre-processing is then applied to each retrieved clip art image for better contrast enhancement. Finally, the pre-processed clip art is displayed in phosphene simulation to mimic the image perception by visual prostheses users.

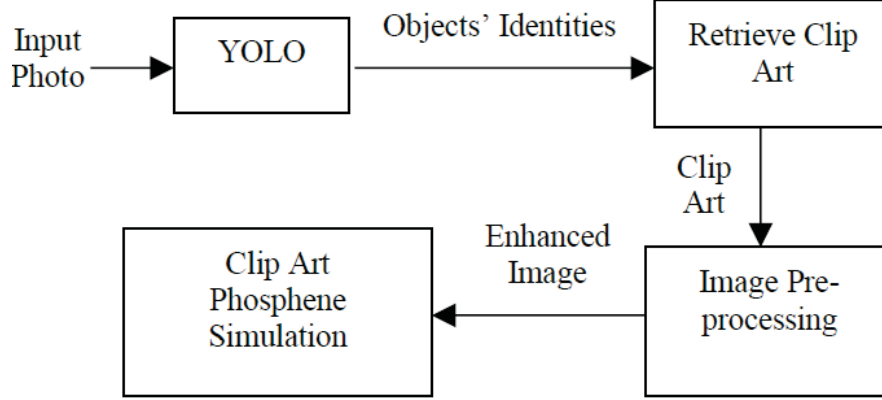


Figure 5.1: Block diagram of the proposed approach.

5.1.2.1 Experimental Design and Procedure

To evaluate the efficacy of the proposed approach, 21 subjects (9 males and 12 females) of age 19 to 65 years (35.05 ± 15.8 years old) participated in three experiments, 7 subjects per experiment. All the subjects in the experiments had normal/corrected vision. None of the participants had a previous experience with simulated prosthetic vision. Using a 15-inch computer screen, simulated prosthetic vision test images were presented to the subjects where the subjects were seated at a distance of 1 meter from the computer screen. This led to a visual field angle of 20° to resemble visual prostheses' visual field [83]. Prior to the actual experiments, each subject was presented with a demonstration to illustrate how the phosphenes look like. In the actual experiments, each subject was presented with 24 test images in each of the three experiments. The duration of displaying each image was 10 sec in which the subjects were given the chance to tell the identity of the objects in the image. The images included objects belonging to various categories such as dog, bicycle, bottle, umbrella ... etc.

The subjects in the first experiment were presented with the test image in its simulated prosthetic vision representation for 3 sec. This is followed by displaying the phosphene simulation of each object in the scene by displaying the object of interest and darkening the rest of the image for 3 sec to accurately localize the objects. The objects are displayed in their order of appearance in the input photo from left to right and from top to bottom. In the last 4 sec, each object was resized to span the full 32×32 resolution (Control Group). This experiment was performed to determine the ability of the subjects to recognize the objects when being resized to occupy the full resolution in the phosphene simulation without

conversion to clip art. In the second experiment, the subjects were presented with only the clip art representation of each object in the photo for the whole 10 sec. In the third and last experiment, the subjects were presented with the same sequence of images that was used in the first experiment. However, the last 4 sec in this case included displaying the clip art representation of each of the objects in their simulated prosthetic vision form, instead of resizing the phosphene simulation of the actual objects.

5.1.2.2 YOLO-based Clip Art Retrieval

YOLO is a deep neural network that has been shown to achieve superior results in object recognition [65]. We utilized YOLO version 3 in the second and third experiments. This version has an architecture of 106 layers (75 Convolutional Neural Network (CNN) and 31 fully connected layers) with skip connections to allow transformation of low-level information. The input high-resolution image (i.e., the actual photo before the pre-processing) is input to YOLO to recognize each of the objects present in the input photo. YOLO recognizes the detected objects, each in a bounding box. Accordingly, we count the total number of detected bounding boxes to identify the total number of recognized objects in the image. The coordinates of the bounding boxes are saved to locate each object in the image.

The labels of the detected images are saved and input one by one to be searched for by Google Images, with the “Clip Art” option automatically selected in Google Images. The first retrieved clip art image is then downloaded for each input label. The file name of the downloaded image is saved and named based on the number of the high-resolution test image followed by the number of the object within that image.

5.1.2.3 Image Pre-processing

To provide better visualization of the images before applying phosphene simulation, we utilized the image enhancement techniques proposed in Chapter 3. In the first and third experiments, if the foreground of the input photo that includes the objects is darker than the background, we use the inversion technique discussed in Chapter 3 so that the foreground is always brighter than the background. This is to guarantee that the objects of interest would be represented by bright phosphenes [102]. We used grayscale phosphene simulation in this chapter unlike the binary phosphene simulation discussed in Chapter 3, to mimic the maximum number of gray levels perceived by a visual prosthetic user [96]. If the foreground is white and the background is black, then no inversion is needed. However, if the foreground is black and the background is white, inversion is performed by subtracting the intensity of each pixel in the grayscale image from 255.

5.1.2.4 Phosphene Simulation

The output of the pre-processing part is utilized to generate a simulated prosthetic vision image. We utilized a square grid for phosphene representation. A circular phosphene shape was also used in this study consistent with multiple studies of prosthetic vision [28]. The distance between each two consecutive phosphenes was set to 0.5 pixel. To map the limited

number of grey levels used in phosphene simulation, only 8 gray levels were used instead of the actual 256 levels (i.e., 0, 36, 72, 108, 144, 180, 216, and 252) [25].

5.1.2.5 Evaluation Metrics

Three metrics were utilized to measure the performance of the subjects in each of the three experiments. First, the time taken by the subject to correctly recognize the presented objects, was recorded, with a maximum of 10 sec. Second, we measured the object recognition accuracy defined by Equation 4.2. Finally, the subjects described their confidence level in their recognition on a scale of 1 to 5. Responses of the participants were logged by the experimenter during the course of the experiment.

5.1.3 Results

5.1.3.1 Object Recognition Enhancement

We first demonstrate the image pre-processing performed prior to phosphene simulation, which utilizes the enhancement technique discussed in Chapter 3 but with grayscale phosphene simulation instead of binary phosphene simulation. We then illustrate the output of each stage of the proposed approach. Figure 5.2 demonstrates an overview of the implemented process showing how clip art representation enhances the object clarification and, thus, object recognition. The input image includes two objects (the bird and the sheep), which belong to two classes from the COCO dataset used in training the YOLO neural network. The figure first demonstrates the phosphene simulation of the entire input image, in which it is extremely hard to recognize that there are two objects. Next, the figure shows the locations of the objects in the photo which illustrates that we have two objects. This allows the subjects to localize the objects within the viewed scene. The figure then shows the two objects after being zoomed in to be occupying the full 32×32 image resolution. Finally, the figure shows the clip art representation of the objects after phosphene simulation. Comparing the phosphene simulation of the clip art representation to that of the original objects demonstrates that the object simplification achieved using clip art representation helps in easily recognizing the identity of the viewed objects.

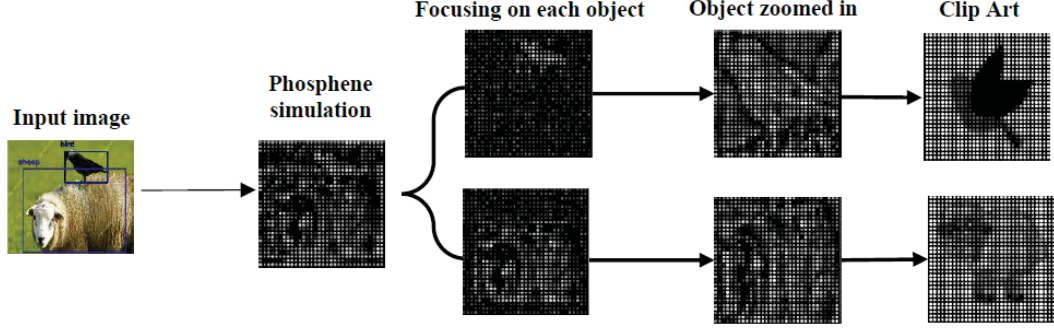


Figure 5.2: Phosphene simulation of each object for actual and clip art representations. The first group of subjects was presented with the phosphene simulation of the input photo, the focused objects and their zoomed in version. The second group of subjects was only presented with the phosphene simulation of the clip art representation only. The third group of subjects was presented with the phosphene simulation of the input photo, the focused objects and their clip art representation.

5.1.3.2 Performance Evaluation

We examined the performance of the proposed approach through experiments that involved 3 groups of subjects, 7 subjects each. The first group of subjects was presented with the phosphene simulation of the photo followed by both the location of the objects in the phosphene simulated image and the resized version of each of object occupying the full 32×32 resolution. The second group of subjects was presented with phosphene simulation of only the clip art representation of the objects. Finally, the third group of subjects was presented with the same sequence of images as that of the first group except for the last image in which the phosphene simulation of the clip art representation of the objects was shown. Figure 5.3a demonstrates the time taken to recognize the objects for each group of subjects. The figure demonstrates a very long duration needed by the first group to recognize the objects that is significantly higher compared to when clip art representation is presented to the second and third groups (Photo Prior to Zoomed in Objects: 9.38 ± 1.31 sec, Clip Art Image Only: 2.73 ± 0.53 sec, Photo Prior to Clip Art Image: 7.32 ± 1.04 sec, $P < 1e - 07$, $n = 24$, two-sample t-test). While the second group of subjects needed the least amount of time to recognize the objects, they were not capable of localizing the objects in the input scene as they were presented with the clip art images directly. However, the third group was capable of localizing the objects given that they were presented with images highlighting the location of the objects prior to displaying the clip art images as shown in Figure 5.2.

We also assessed the ability of the test subjects in each group to correctly recognize the presented objects. Figure 5.3b demonstrates the recognition accuracy, where the first group of subjects achieved the least recognition accuracy. On the other hand, the recognition accuracy for both the second and third groups of subjects was identical due to relying on clip art in both experiments. In this case, the subjects' answers were identical as they were all able to recognize the objects correctly in their clip art representation (Photo Prior to Zoomed in Objects: $28.82 \pm 31.04\%$, Clip Art Image Only: $100 \pm 0\%$, Photo Prior to Clip Art Image: $100 \pm 0\%$, $P < 1e - 04$, $n = 24$, two-sample t-test).

We finally assessed the confidence of the test subjects of each group in their recognition of

the presented objects. Consistent with the aforementioned results, Figure 5.3c demonstrates that the confidence level of the participants was significantly high when they were presented with the clip art representation of the objects (i.e. the second and third groups of subjects), unlike that of the first group (Photo Prior to Zoomed in Objects: 2.48 ± 0.94 , Clip Art Image Only: 4.99 ± 0.06 , Photo Prior to Clip Art Image: 4.96 ± 0.12 , $P < 0.05$, $n = 24$, two-sample t-test).

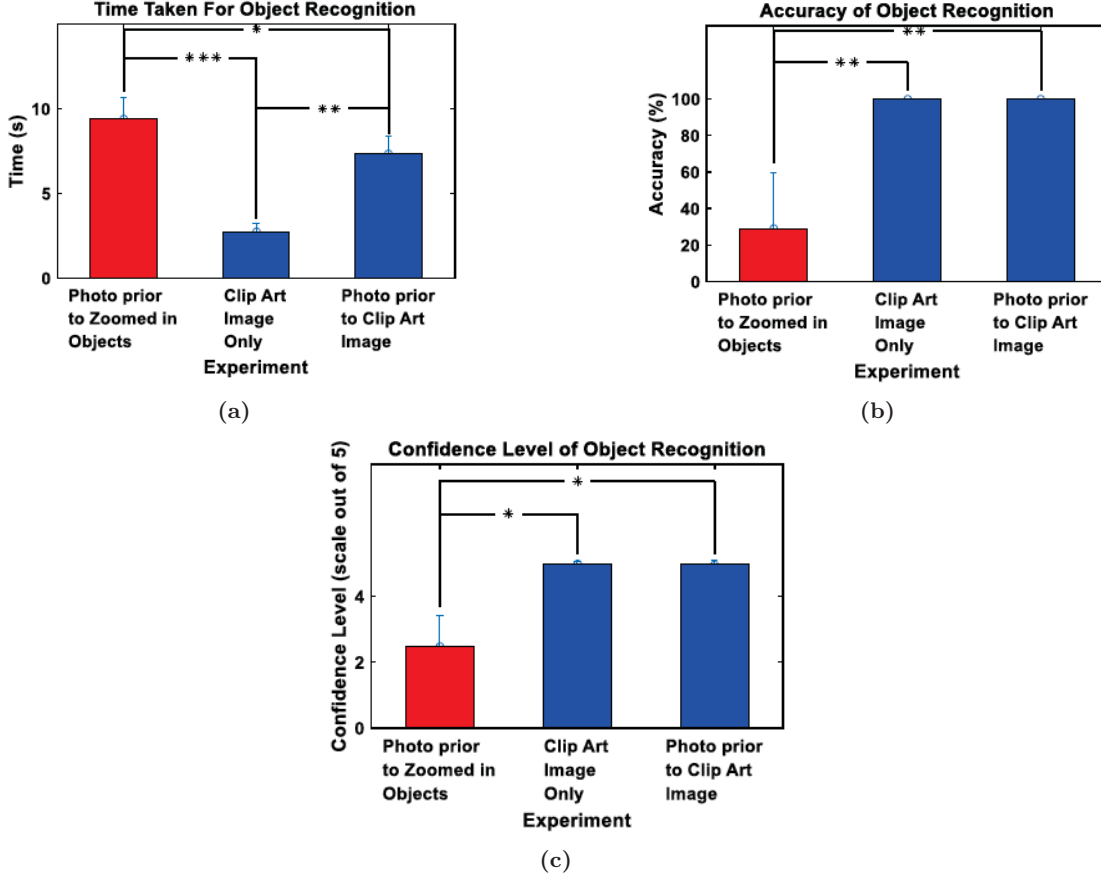


Figure 5.3: Performance of different groups of subjects in the task of recognizing the objects measured by (a) time to decision, (b) recognition accuracy, and (c) confidence level (*mean \pm std*). Blue bars represent using clip art for object simplification, while the red bar represent zooming on the actual object of interest. $*P < 0.05$, $**P < 1e - 04$, $***P < 1e - 07$, two-sample t-test.

5.1.4 Conclusion

Visual prostheses have recently provided hope to the blind. However, a core challenge that affects the perception of visual prostheses devices is the difficulty of object recognition due to the low resolution of the perceived image. We introduced an automated approach to enhance the process of object recognition. This was achieved by first using YOLO deep neural network to detect and recognize the objects in a scene. The corresponding simplified clip art representation is then obtained and displayed to the user. Three experiments were performed to evaluate the performance of the proposed system, using phosphene simulation, where one experiment relied on only the input photo without using the clip art. The other two experiments relied on clip art representation. The results demonstrate that using the clip art representation of the objects outperforms using the actual image in terms of recognition time, recognition accuracy and confidence level.

5.2 A GAN Application for Clip Art Generation

5.2.1 Introduction

Recently, deep learning, and GANs in particular, have been employed in different tasks that involve image-to-image translation [103]. Generative modeling is an unsupervised learning task in machine learning which automatically discovers and learns patterns in input data so that the model can be utilized to generate new examples that look similar to a given original dataset [104]. GANs consist of two main components: a generator and a discriminator. The generator generates candidate synthetic outputs, while the discriminator evaluates them [105]. The main aim of the generator is to try to fool the discriminator by producing novel candidates that are indistinguishable from the real data. Backpropagation is applied to both networks so that the generator produces better samples, whereas the discriminator becomes more skilled at figuring out fake samples [69].

GANs have been frequently used in image-to-image translation tasks such as transforming an image to its winter version, a day-light image to its night version and a real image to its corresponding cartoon version [106–108]. For instance, one of the available GANs, StackGAN, generates photo-realistic images from text interpretations. It is composed of two stages: one stage takes the text interpretation as input and sketches the primitive shape and color of the object accordingly. The second stage takes as input the results from stage one along with the text descriptions to generate a high-resolution image with realistic details [109]. GANs have also been used in generating pairs of corresponding images in two different domains known as coupled GAN. The coupled GAN framework is decomposed of two sub-networks: one network is responsible for low-resolution images and the other network is responsible for high-resolution images where each sub-network tries to maximize the pair-wise correlation between the two feature domains in an ordinary embedding subspace [110]. Additionally, a well-known GAN model is conditional GAN that aims to generate samples from distributions satisfying certain conditioning on some correlated features. Regular GAN and conditional GAN differ primarily in that conditional GAN has control over the modes of the data generation, whereas regular GAN does not [111]. Conditional GAN can be performed on discrete labels

which are discrete values given for each class [74, 112, 113], text [114] and images. The image-conditional models have been utilized in image prediction from a surface normal map [115], product photo generation [116], future frame prediction [117] and image generation from sparse annotations [118–120]. Multiple studies have also developed GANs for image-to-image mappings by applying the GAN unconditionally, relying on other terms, such as L2 regularization, to force the output to be conditioned on the input. These studies demonstrated outstanding results on future state prediction [121], inpainting [122], image manipulation guided by user constraints [123], super-resolution [124] and style transfer [125]. One GAN model that has demonstrated success in multiple applications is Pix2Pix GAN which uses conditional GANs to perform image-to-image translation. In this process, one image belonging to a certain domain is entered along with the corresponding image from another domain, both combined as one image, with the objective of learning the mapping between the two domains [126]. Another model is CycleGAN which uses unpaired images from different domains and attempts to learn the mapping that was done in the paired image-to-image translation [72].

We propose ClipArtGAN, a GAN that performs novel adjustments to the Pix2Pix GAN input data to generate clip art images of objects represented in photos. To our knowledge, this is the first attempt to develop a GAN model for clip art generation. The adjustments performed to the clip art images are for their color and pose, based on the corresponding photo. Thus, the generated clip art image possesses the same orientation of the real object in the photo and will be given a color similar to that of the object. Evaluating the proposed ClipArtGAN through multiple experiments demonstrate the ability of this approach to generate representative clip arts.

We utilized the ClipArtGAN model proposed in this thesis to enhance the ability of visual prostheses’ implanted patients to recognize the perceived objects. This proposed application was called PVGAN hereafter. Given the low resolution of the perceived prosthetic image, we propose a novel deep learning approach for object simplification to enhance the ability of visual prostheses’ users in object recognition.

In this thesis, we propose PVGAN; a model trained using images of objects belonging to general classes that visual prostheses users could encounter in their daily life. The model performs novel adjustments to Pix2Pix input data to generate clip art images from a real high-resolution image. The generated clip art image is then used as input to the visual prosthesis system to enhance object recognition for implanted patients. The proposed approach is evaluated using simulated prosthetic vision experiments that involve normally sighted participants demonstrating the efficacy of the proposed approach.

5.2.2 PVGAN Model Overview

PVGAN uses the proposed ClipArtGAN architecture but the only difference is in the used number and nature of the training classes. ClipArtGAN can be used in any field however, in PVGAN, the training classes of ClipArtGAN are updated to fit in the visual prostheses field; thus, the name PVGAN (Prosthetic Vision Generative Adversarial Network). The proposed model uses a GAN to generate clip art from input images. Figure 5.4a shows the training process of the proposed model architecture, where an input image that consists of a high resolution image, referred to as photo hereafter, concatenated with its corresponding

clip art image is entered to the generator. The generator then attempts to generate a clip art that is similar to that of the provided clip art of the input image to be able to fool the discriminator. On the other hand, the discriminator takes as input the original clip art of the input image in addition to the generated clip art image to decide whether the generated clip art by the generator is indistinguishable from the original clip art or not. The generator keeps improving its ability in generating clip art through training until the discriminator cannot differentiate between the original and the generated clip art images. Once the model is trained, the generator of the trained PVGAN is used to generate a clip art that corresponds to an input photo. The object in the input photo is not restricted to the objects that were used in training the model. For instance, the trained PVGAN can be used to generate a clip art that corresponds to a sofa (as shown in Figure 5.4b) without previous training on photos and clip arts of sofas. Thus, the proposed PVGAN learns to generalize the generation of clip arts from photos of arbitrary objects. While the PVGAN model proposed in this thesis uses the same architecture as that of the Pix2Pix GAN, manipulating the training dataset has a significant impact on the ability of the proposed GAN to learn the intended mapping between high-resolution photos and clip arts. Finally, the PVGAN model generates clip art image that will be used as an input to a phosphene simulation procedure to produce an image that contains phosphenes matching the resolution of visual prostheses to simulate the image perceived by visual prostheses users.

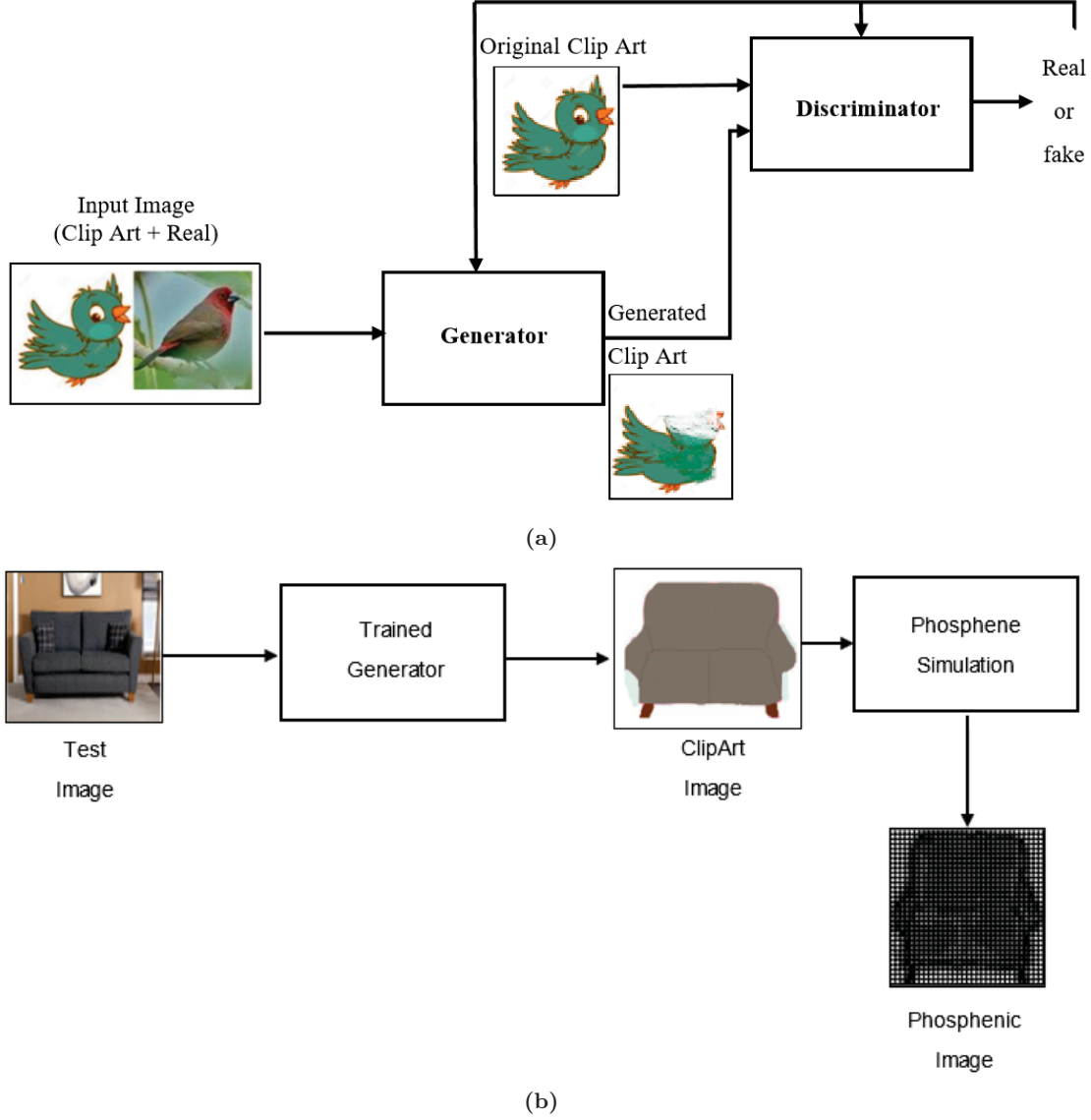


Figure 5.4: Block Diagram showing the flow of the proposed model. (a) ClipArtGAN/PVGAN Training. (b) Testing the Model and Phosphene Simulation.

5.2.3 Training Datasets

Preparing the dataset used in this thesis was a challenge in many aspects to produce the best possible results when used to train our PVGAN model. The dataset used in this study was collected from various sources [127–131] in addition to Kaggle website [132–138]. The dataset used in ClipArtGAN comprises 12 classes: apple, banana, bike, bird, car, dog, elephant, flower, horse, lemon, pizza and tree. On the other hand, PVGAN utilized a dataset that is composed of images belonging to 17 classes representing food (apple, banana, lemon and pizza), vehicles (bus and car), animals (bird, dog and zebra), furniture (chair, bed, door and table), plants (flower and tree) and personal belongings (bag and laptop). These classes were selected to ensure diversity across the objects used to train the model. The dataset included only real high-resolution images (i.e., photos). In order to prepare the dataset, the following steps were performed. First, we included 175 photos per class that show only one object of

interest centered in the middle of the image. Second, we renamed the photo files based on the class the image belongs to (e.g., “elephant1.jpg”, “bird99.jpg”, ... etc.). Thus, the total number of the photos that we used for training our PVGAN model is 2100 images. To obtain the clip art corresponding to the photos of each class, we used Google Image search. Three different datasets of clip art images were formed to train the GAN models. The first training dataset, named as “First Choice” (FC) hereafter, was obtained by iterating on the 2100 photos and parsing the name of each file until reaching a digit. The word preceding the digit, such as “elephant” in “elephant1.jpg”, was used to search Google Images after selecting the Clip Art in the type of the retrieved image. In the FC dataset, the first output clip art returned through the search was used as the corresponding clip art. We then concatenate each clip art image with its corresponding photo to be ready for training the model.

The second training dataset, named hereafter as “First Choice-Orientation and Color Matching” (FC-OCM) used the same images used in FC dataset but with orientation and color adjustment based on the input photo. To adjust the pose of the clip art object, the pose of the object in the high-resolution image as well as in the corresponding clip art need to be identified. This is done by, first, applying a median filter of size 5×5 to remove any noise from the background of the photo [139]. Second, we extract the object of interest using a bounding box by means of connected components labelling using 8-connectivity [140] and place it on another image with a black background and no objects. Third, we binarize the new photo (i.e., the black image with the object of interest extracted on it) by applying Otsu thresholding which is defined as finding the threshold value where the difference between the foreground and the background is maximized to be ready for applying the skeletonization morphological operator [141]. Skeletonization is performed by means of hit-or-miss transform to maintain the shape topology [142]. The structuring element size used in the skeletonization operation was 3×3 . After skeletonization, pixel-wise gradient orientation of the new photo is calculated using Prewitt operator to get the directions of the edges in the object of interest [143]. Finally, the average of all gradient orientations of the skeleton is computed and used as an estimate of the orientation (pose) of the object. The same procedure is also applied to the clip art image to get the average of the gradient orientations of its skeleton. To match the pose of the clip art object to its high-resolution counterpart, we subtract the average of the gradient orientations of the clip art object from that of the object in the photo. We, then, rotate the clip art object based on that difference to get the same orientation as that of the object in the photo. We also matched the color of the clip art object to that of the corresponding high-resolution object. This is done by, first, computing the average color of the object of interest in the photo. Second, the average color of the object in the clip art image is calculated.

The difference between both averages is then obtained and added to the color of each pixel of the clip art object. Clipping was applied to ensure that the new colors are within the range of colors (i.e., $[0, 255]$ for each of the red, green and blue components of each pixel). The third training dataset, named hereafter as “Histogram of Oriented Gradients-Orientation and Color Matching (HOG-OCM)”, adjusted the orientation and the color, same as in FC-OCM, but with a different criterion of clip art choice. In this dataset, instead of taking the first-choice clip art image retrieved in the search result, the first 10 clip art images are downloaded to increase

the probability of having a clip art image that is similar in shape to that of the original object in the photo. The gradient magnitude and direction are computed for the input photo and each of the 10 downloaded clip art images as

$$GradientMagnitude = \sqrt{(g_x)^2 + (g_y)^2} \quad (5.1)$$

where g_x and g_y are the gradients in the x and y directions, respectively. The direction is expressed as

$$GradientDirection(\phi) = \tan^{-1} \left(\frac{g_y}{g_x} \right) \quad (5.2)$$

Unlike the FC and FC-OCM datasets, downloading the first 10 clip art images provides a variety of choices to select the clip art image that best matches the photo. To perform this selection, the Histogram of Oriented Gradient (HOG) of each of the high-resolution photo and each of the 10 clip art images retrieved through Google Image search was computed. The features obtained by HOG encode local shape information [144]. HOG was used with a cell size of 8×8 to maintain the small-scale details with nine orientations histogram bins which are 0, 45, 90, 135, 180, 225, 270, 315 and 360. Moreover, the block size used in the HOG operation was 2×2 to help suppress illumination changes of HOG features [145]. Next, Euclidean distance was calculated between the HOG feature vector obtained for each of the 10 clip art images and that of the high-resolution photo. The length of the HOG feature vector per block is 36 calculated by multiplying the block size by the number of bins (i.e., $2 \times 2 \times 9 = 36$). The images' size in the dataset was all resized to be 256×256 to match the standard size used in Pix2pix GAN. Therefore, the feature vector per image is the total number of blocks $\times 36$ (i.e., $16 \times 16 \times 36 = 9216$). The best matching clip art image was determined based on the minimum Euclidean distance obtained.

5.2.4 Pre-processing of Training Input Data

Prior to training the model, pre-processing for the input dataset is performed. First, normalization of the images is performed by rescaling the pixel value to match the standard rescaling factor used in Pix2Pix GAN expressed as [75]

$$p = \frac{p}{127.5} - 1 \quad (5.3)$$

where p is the pixel intensity that is being rescaled by the middle value of the intensities. Second, all images were resized to the same size (i.e., 256×256). Finally, random flipping is performed to increase the generalization of the model where horizontal flipping is used in this case. Based on this step, the total number of training images is 4200 images. In order to match the input images to the required input of Pix2Pix, the input images were entered as pairs of images stitched together where the left image is the clip art version of the right high-resolution image. All the photos consisted of only one object of interest to train the model more accurately.

5.2.5 Phosphene Simulation

The generated clip art image is further processed to be expressed in terms of phosphenes to simulate the prosthetic vision. The grid used in the simulation is a squared grid [28, 91]. The limited number of electrodes in visual prostheses limits the number of pixels perceived by visual prostheses users. To accommodate for the number of electrodes available in the implants based on the developments of the number of electrodes used in future versions of the Argus device, Alpha IMS and Polyretina, a square lattice of size 32×32 pixels was used [25, 97]. The brightness of the phosphenes changes as a function of stimulation intensity across electrodes, thus, the brightness of phosphenes elicited by an individual electrode should scale appropriately with luminance [146]. We used 8 gray levels in our simulations to match the mid-range of the values reported in the literature [96]. Two types of phosphene simulation models were used to assess the effect of using the generated clip art from PVGAN; namely scoreboard and axon map models.

5.2.5.1 The Scoreboard Model

In this model, the phosphenes are of an idealized circular shape based on the common form of phosphenes reported in literature [28, 91]. Using this pixel-based approach, a visual image is recreated assembling phosphenes into objects and images similar to an electronic scoreboard [18]. The distance between each two consecutive phosphenes was set to 0.5 [90]. The 8 gray levels used are 0, 36, 72, 108, 144, 180, 216 and 252 to uniformly map the range of gray levels from 0 to 255.

5.2.5.2 The Axon Map Model

Recent studies have reported that phosphene shapes, especially in epiretinal implants, have some spatial and temporal properties resulting in distortions and temporal fading [147]. This distorted shape could follow the shape of the axons sent by the retinal ganglion cells; hence, the name axon map. Therefore, we also examined this model to assess the efficacy of the proposed PVGAN when applied to this type of phosphene simulation. In this model, we used the pulse2percept Python library using 8 gray levels [31, 148].

5.2.6 ClipArtGAN Evaluation Experiments

Prior to the utilization of ClipArtGAN in visual prostheses, we performed a number of experiments to ensure the efficacy of the model in terms of correctly generating the clip art for a certain high-resolution image. Three experiments were conducted to evaluate the ability of ClipArtGAN and each of the training datasets to generate representative clip art images. In the first experiment, named “CycleGAN vs Pix2Pix” hereafter, the clip art images generated from CycleGAN and Pix2Pix models were compared. Such comparison was performed after training each of the two models with each of the three training datasets (FC, FC-OCM and HOG-OCM). Twelve subjects (9 females and 3 males) participated in the experiment in which their task was to recognize the object from the generated clip art. The experiment contained two sets of images; each consisting of 12 images. The first set includes 12 images; one from each of the 12 classes considered in training the models. However, these images are different

from the images used in the training process. The second set involved 12 images that belong to classes different from the ones used in training; namely bear, broccoli, orange, tomato, cat, zebra, clock and moon. Each class from the new classes that belong to the second set includes 10 images so the total number of test images available in the second set was 80 images. This experiment was performed to be able to know the efficiency of each of the three training datasets in training the models (i.e., CycleGAN and Pix2Pix GAN) in terms of measuring the recognition accuracy. The accuracy was computed for each participant as the number of clip art images correctly recognized by the participant divided by the total number of images presented. All participants were able to notice that the image displayed in front of them is a clip art image, so the only metric measured in this experiment was the ability of the participants to recognize the objects correctly.

The second experiment, named “Pix2Pix GAN vs Google Images” hereafter, involved the same 12 participants where this time they were presented with the generated image from the three training datasets using Pix2Pix against the clip art result for the same test image in Google Images. The participants were asked whether the generated clip art image from the model or the retrieved clip art image from Google Images is a better representation of clip art that corresponds to the object in the photo. The clip art images collected by Google Images were retrieved by uploading the test image to Google Images choosing the clip art option to get the clip art of the uploaded image. The recognition accuracy was based on whether the generated or retrieved image is a clip art or not.

The third experiment, named “Google Images Recognition” hereafter, was performed using Google Images without the need for any participant. In this evaluation, the 24 generated clip art images that correspond to the 24 test images obtained using CycleGAN and Pix2Pix, were uploaded through Google Images to measure the ability of Google Images to recognize the object in the generated clip art. This was done by saving the objects’ names of each of the 24 generated clip art images in an array and saving the corresponding generated ClipArt images in another array. Then, we upload each of the generated clip art images to Google Images and retrieve the most similar visual images that appears in the Google search. Next, we search using the corresponding name of the object for the same name in the results obtained from Google Images. This is performed by parsing the search bar that includes the description provided by Google Images for the uploaded image search of the object’s name. If there is a match, this implies that Google Images was able to recognize the clip art image generated using the corresponding model. In this case, the recognition accuracy is computed as a ratio of correctly recognized test images relative to the total test images.

5.2.7 PVGAN Experiments Overview and Setup

To ensure the efficacy of the proposed PVGAN model that is used in visual prostheses, a number of experiments was conducted. In these computer-screen experiments, different metrics were measured to evaluate the performance of the subjects when presented with phosphene simulations of the generated clip art images from the PVGAN model. A set of 24 generated clip art images displayed after phosphene simulation was used to simulate the perception of images perceived by any visual prostheses implanted patients. The set was divided evenly into two equal groups (i.e., each group is of 12 test images), where one group has new images from

the training classes that the PVGAN was trained on, and the other group has test images from new classes that the PVGAN model did not see before. All subjects had no prior experience with simulated prosthetic vision. The subjects did not receive any payment for participating in the study.

5.2.7.1 Scoreboard Phosphene Simulation Experiments

The first set of experiments involves four groups of subjects which were conducted using the scoreboard phosphene simulation model, where each experiment comprised different subjects. Each of the four experiments involved different subjects to avoid the learning effect that might occur if the subjects were presented with different versions of the same image. The first group, named as “Real Image” hereafter, involved 10 subjects, 7 females and 3 males aging from 22-60 years old (31 ± 11.47 years old). In this group, phosphene simulation of each image from the 24 images was displayed for 10 seconds with automatic switching between successive images (Control Group). The second group, named as “Real before Clip Art” hereafter, involved 10 different subjects. The subjects were 5 females and 5 males aging from 27-55 years old (36.5 ± 10.24 years old). In this group, phosphene simulation of the real image (i.e., photo) was displayed for 5 seconds followed by phosphene simulation of the corresponding generated clip art image displayed for another 5 seconds. So, a total duration of 10 sec per each real image with its corresponding generated clip art image was used during this experiment to match the same time per image used in the “Real Image” group. The third group, named as “Clip Art no Dropout” hereafter, involved 10 different subjects; 6 females and 4 males aging from 24-40 years old (32 ± 6.2 years old). In this group, only the generated clip art images were displayed using phosphene simulation, each for 10 seconds to match the duration taken per one version of an image in both the “Real Image” and the “Real before Clip Art” groups. Finally, the fourth group, named as “Clip Art with Dropout” hereafter, involved 10 different subjects; 8 females and 2 males aging from 20-60 years old (37.7 ± 15.64 years old). In this group, the generated clip art images were displayed with dropout added to the phosphene simulation to test whether the subject will be still able to recognize the object after the clip art generation or the dropout will hinder the recognition. Each of the images was displayed for 10 seconds to match the duration used in all the other three experiments per one image. The dropout percentage used was 20%, where the dropout locations were randomly assigned [149].

5.2.7.2 Axon Map Phosphene Simulation Experiments

The axon map model was also used to take into account the spatial and temporal fading and distortions that happen to the phosphenes. The same set of the 24 test images used in the scoreboard phosphene simulation experiments were used in this experiment with a different set of subjects. In these experiments, a total of 10 subjects were involved in this experiment, 5 males and 5 females aging from 19-57 years old (34.1 ± 13.17 years old). The subjects in this experiment were different from those who participated in the other four experiments to avoid any learning effect since the same set of images (i.e., themes) were used in this experiment. In contrast to the scoreboard model experiments, this experiment comprised the same set of 10 subjects who were presented with 24 images that belong to four types, 6 images per types. These types match the groups used in the scoreboard experiments (i.e., Real Image, Clip Art

no Dropout, Real before Clip Art, and Clip Art with Dropout). The order of images displayed throughout the experiment was randomly shuffled across the subjects. The dropout used in the axon map phosphene simulation experiment was added manually after image generation as the pulse2percept library used, for this axon map simulation, does not implement the dropout [31].

5.2.8 PVGAN Experiments Evaluation Metrics

The subjects in all of the conducted experiments were asked to try to recognize the objects appearing in front of them. Responses of the subjects were logged by the experimenter during the course of the experiment. Three evaluation metrics were used to evaluate the performance of the subjects in each experiment. The metrics are the time taken to recognize the object measured in seconds, the recognition accuracy measured as the recognition score averaged across images and subjects, and the confidence level on a scale from 1 to 5, where 1 indicates that the subject is extremely unsure about the answer, whereas 5 indicates the contrary. The recognition score for one subject per image was either 0% for incorrectly recognizing the object or not being able to decide, 100% for recognizing the object correctly and 50% for recognizing the general theme of the object but incorrectly recognizing the object itself. For example, an accuracy of 0.5 was given if the subject recognizes a couch as a chair given that both have the same general theme. Subjects were informed in advance that there will be 24 images in the experiment and they were informed about the available themes of the objects.

5.3 Results

5.3.1 Training Datasets Pre-Processing Overview

We first demonstrate the outcome of different pre-processing steps performed to the clip art images in the training datasets. Figure 5.5 demonstrates the steps performed to align the downloaded clip art image with its corresponding photo in terms of both the color and the orientation (i.e., pose) of the photo. First, a bounding box that surrounds the object of interest in the photo and the downloaded clip art image is extracted. Second, the extracted object of interest is binarized using Otsu thresholding to give the object of interest a white color and the background the black color to be ready for the skeletonization step. Third, by means of hit-or-miss morphological operators, skeletonization is applied to both the input photo and its corresponding clip art image. The difference between the average colors of the extracted object of interest in both the photo and the clip art image is calculated. This difference is then used to update the colors of the object of interest in the clip art image. Finally, the clip art image is adjusted in terms of both its color and pose according to that of the photo.

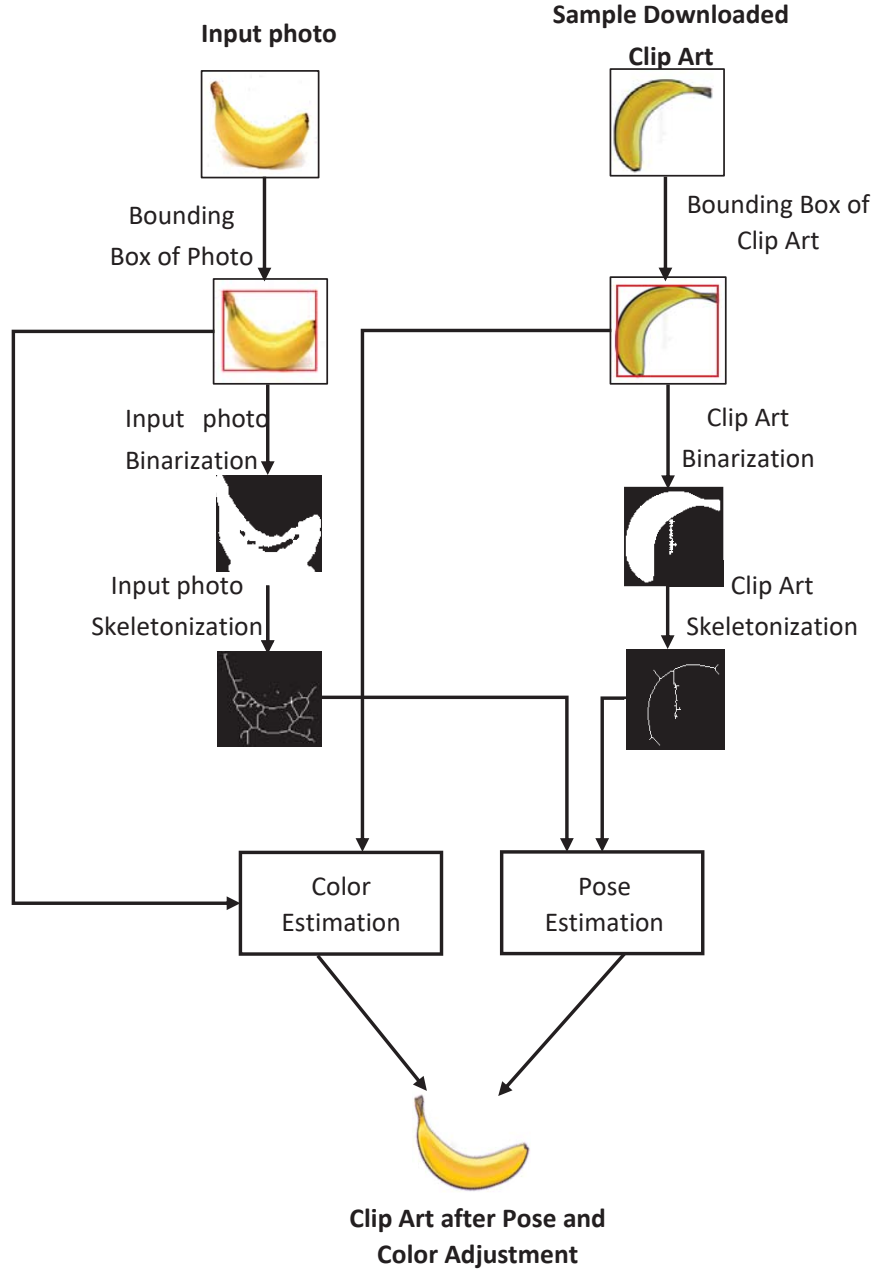


Figure 5.5: Clip Art pre-processing.

We then demonstrate the clip art selection process using HOG that was used to form the HOG-OCM training dataset. Figure 5.6 demonstrates the criterion used to select the best matching clip art image out of the first 10 downloaded images that is used in the HOG-OCM training dataset. A sample photo is taken as input to HOG to visualize the gradient directions for every 8×8 pixel in each 2×2 cell within the 256×256 image. In addition, HOG was applied to three sample clip art images from the 10 downloaded images where the gradient directions are shown on top of each of the three clip art images. Moreover, for better visualization, zooming in certain part in each of the images is shown to show the orientation of the 9 bins of the histogram. The best matching clip art is obtained by computing the Euclidean

distance between the HOG features of the input photo and those of the three clip art images. As shown in the figure, the selected clip art has the most similar shape and orientation to that of the input photo among the three clip art images.

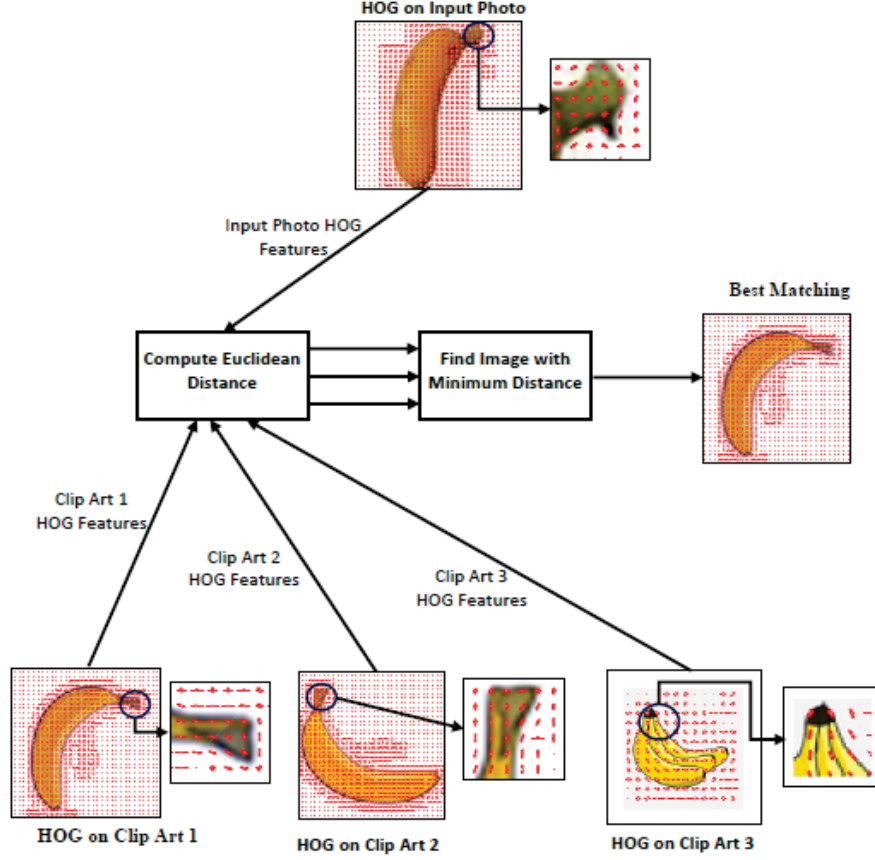


















Figure 5.6: HOG best matching clip art selection.

5.3.2 Training Datasets Qualitative Evaluation







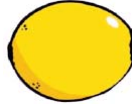









We next examine samples of the outputs of each of CycleGAN and Pix2Pix, shown in Table 5.1 and Table 5.2, respectively, when trained using each of the three training datasets (FC, FC-OCM and HOG-OCM). The tables demonstrate the output of each of CycleGAN and Pix2Pix when tested using images belonging to the same 12 classes of the training dataset; however, these specific images were not part of the training dataset. In addition, the tables also demonstrate the output of both GANs for the more general case when tested using images that do not belong to the 12 classes of the training dataset. Table 5.1 shows that for simple test images such as the lemon and the orange, CycleGAN was able to generate clip art images which are circular or spherical in shape similar to the input objects. However, for the lemon, it failed to generate a representative clip art, but instead adopted the overall shape of an apple. In addition, for more complex images, CycleGAN failed to generate representative clip art as in the example of the dog. Comparing the performance across the three training datasets FC, FC-OCM and HOG-OCM does not reveal significant difference. However, the outcome of HOG-OCM could be considered more similar to the original object.

Table 5.1: Testing the CycleGAN model.

	Test Image	CycleGAN Predicted Image		
		FC Dataset	FC-OCM Dataset	HOG-OCM Dataset
New images belonging to the same classes encountered in training				
				
Images from new classes not encountered during training				
				

Using Pix2Pix resulted in a much better output compared to that of CycleGAN. Table 5.2 reveals that for the new test images that belong to classes encountered during training, representative clip art that matches the identity of the given test image is generated. This can be observed for the dog and lemon sample images for all three training datasets. However, comparing the three training datasets, the clip art image generated from using the HOG-OCM dataset in training the Pix2Pix GAN was better than that of the other two datasets since the shape of the clip art generated matched that of the test image. More importantly, representative clip art is also generated for new classes not encountered during training as in the example of the bear and the orange. For the bear test image, the output of Pix2Pix trained using FC and FC-OCM datasets adapted the elephant image that was present in the training datasets to match the form of the bear. This could be explained since the elephant is the only class in the training classes that have the same body appearance of the bear (i.e., a big animal).

Table 5.2: Testing the Pix2Pix model.

	Test Image	Pix2Pix GAN Predicted Image		
		FC Dataset	FC-OCM Dataset	HOG-OCM Dataset
New images belonging to the same classes encountered in training				
				
Images from new classes not encountered during training				
				

The same result could be observed for the example of the orange. These results demonstrate that the clip art images generated using Pix2Pix trained with the HOG-OCM dataset results in the best performance, which is significantly better than the clip art images generated using CycleGAN. As a result, we refer to Pix2Pix trained using the HOG-OCM dataset as ClipArtGAN hereafter and the proposed GAN utilized in visual prosthesis as PVGAN hereafter.

Given that clip art images could be obtained by searching via Google Images, we compare the output obtained using ClipArtGAN to that obtained through a simple search via Google Images. Figure 5.7 shows a sample of two test images used in Google Images to get their corresponding clip art compared to the generated clip art image obtained using ClipArtGAN. The figure demonstrates that the image resulting from Google Images search seem like a real image not a clip art, although the “ClipArt” option was chosen. On the other hand, clip art generated using ClipArtGAN is a better clip art representation of the test image.

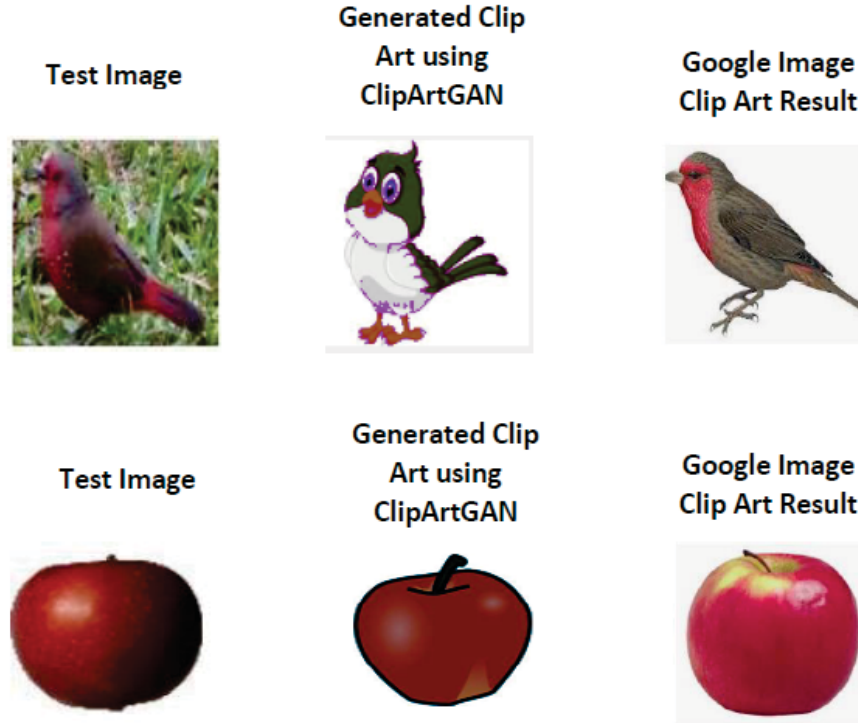


Figure 5.7: Generated Clip Art versus Google Images Clip Art.

5.3.3 ClipArtGAN Experimental Evaluation

We next aimed to quantify the performance of the proposed approach through three evaluation experiments. In the first experiment, the performance of CycleGAN and Pix2Pix was evaluated by participants, where we assessed the ability of the participants to recognize the generated clip art using both models trained using each of the three training datasets (FC, FC-OCM, and HOG-OCM). Table 5.3 shows the average of recognition accuracy of the participants for both CycleGAN and Pix2Pix with respect to the three training datasets. The table demonstrates a diminished recognition accuracy using all training datasets for CycleGAN with a highest recognition accuracy achieved using the HOG-OCM training dataset of 27% and 22% for images belonging to the training classes and images belonging to new classes, respectively. On the other hand, Pix2Pix GAN shows significantly better results compared to that of CycleGAN, achieving a highest recognition accuracy achieved using the HOG-OCM training dataset of 97% and 95% for images belonging to the training classes and images belonging to new classes, respectively. The standard deviations of the recognition accuracy for all training datasets used in both CycleGAN and Pix2Pix show that all the participants were having consistent accuracies. For both GANs, using the HOG-OCM dataset achieves better recognition accuracy compared to using the FC and FC-OCM datasets.

In the second experiment, the performance of Pix2Pix trained using each of the three datasets was compared to the output of Google Images search. The generated clip art images were presented to the participants and their feedback about whether the images represent a clip art or not was quantified. Table 5.4 shows the recognition accuracy achieved comparing Pix2Pix GAN with Google Images for clip art generation. Given the poor performance of CycleGAN in the first experiment, we focused in the second experiment on Pix2Pix only. The

results are shown in Table 4 where it compares the performance of the three training datasets using Pix2Pix GAN with the results obtained for the same test images in Google Images. The resulted images from the clip art search in Google Images were images that are visually similar to the test image that appeared more of photos than being actual clip arts. Thus, the participants’ recognition accuracy reported in Table 5.4 reflects their ability to identify which of the two methods (Pix2Pix or Google Images) is a better clip art representation of the test image. The images generated using Pix2Pix GAN trained using the HOG-OCM dataset resulted in a recognition accuracy of 96% which is significantly higher than the 42% accuracy that is achieved using Google Images search.

As a last method of validation, we performed a third experiment that did not involve participants. In this experiment, the images generated by both CycleGAN and Pix2Pix trained using the three datasets were uploaded to Google Images. The ability of Google Images to recognize the objects in the generated clip arts was quantified. Table 5.5 shows the results confirming the superiority of Pix2Pix GAN combined with the HOG-OCM training dataset in comparison to other models achieving an accuracy of 95% and 89% for images belonging to the training classes and images belonging to new classes, respectively.

Table 5.3: CycleGAN vs Pix2Pix GAN for ClipArt recognition Experiment.

Model	Dataset Version	Accuracy of new images from training classes (<i>mean</i> \pm <i>std</i>)	Accuracy of new images from new classes (<i>mean</i> \pm <i>std</i>)
CycleGAN	FC-Dataset	$20 \pm 6.79\%$	$17 \pm 6.13\%$
	FC-OCM Dataset	$25 \pm 7.15\%$	$21 \pm 6.94\%$
	HOG-OCM Dataset	$27 \pm 7.65\%$	$22 \pm 6.87\%$
Pix2Pix GAN	FC-Dataset	$80 \pm 3.24\%$	$51 \pm 7.57\%$
	FC-OCM Dataset	$85 \pm 2.43\%$	$70 \pm 5.45\%$
	HOG-OCM Dataset	$97 \pm 1.15\%$	$95 \pm 1.27\%$

Table 5.4: Pix2Pix GAN vs Google Images for ClipArt generation Experiment.

Model	Dataset Version	Accuracy of generated images from the model (<i>mean \pm std</i>)	Accuracy of the images shown from Google Images (<i>mean \pm std</i>)
Pix2Pix GAN	FC-Dataset	$65 \pm 6.27\%$	$42 \pm 6.93\%$
	FC-OCM Dataset	$77 \pm 4.73\%$	
	HOG-OCM Dataset	$96 \pm 1.13\%$	

Table 5.5: Google Images recognizing generated Clip Art Experiment.

Model	Dataset Version	Accuracy of Google Images in recognizing new images from training classes	Accuracy of Google Images in recognizing new images from new classes
CycleGAN	FC-Dataset	23%	19%
	FC-OCM Dataset	20%	17%
	HOG-OCM Dataset	25%	21%
Pix2Pix GAN	FC-Dataset	82%	47%
	FC-OCM Dataset	83%	71%
	HOG-OCM Dataset	95%	89%

5.3.4 Phosphenes Simulation Outcome

Following the success of ClipArtGAN in generating clip art images for any high-resolution image, we retrained the model to include the classes that a typical visual prostheses user will deal with in his/her daily life, where we named the newly trained network as PVGAN. We first

demonstrate the outcome of phosphene simulation using both models examined; namely, the scoreboard model and the axon map model. Figure 5.8 shows that the shape of the phosphenes in the scoreboard simulation is a circle-like shape, whereas that of the axon map phosphene simulation has a tail-like shape (i.e., elongated with distortions). In both cases, a square grid of size 32×32 phosphenes was used with 8 gray levels.

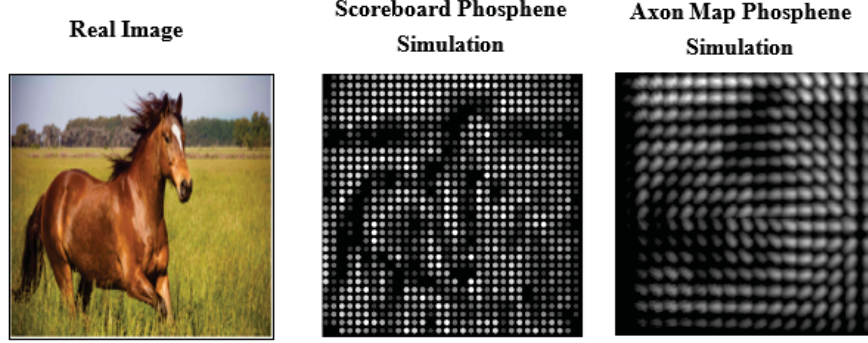

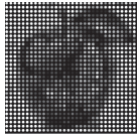


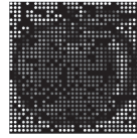




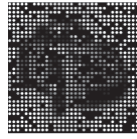


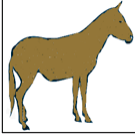

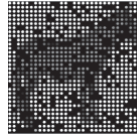

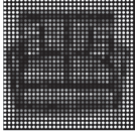

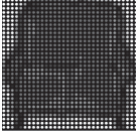
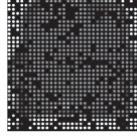


Figure 5.8: Scoreboard versus axon map phosphene simulation.

We next demonstrate the outcome of PVGAN when applied to sample images. Table 5.6 shows the outcome of applying PVGAN to a sample of new test images that either belong to the training classes or belong to new classes. It also shows the scoreboard phosphenes simulation of the real high-resolution image in addition to that of the corresponding clip art version generated from the generator of the PVGAN model. Two sample images are shown that belong to the training classes representing new images other than the images that the PVGAN model was trained on (an apple and a car). The last two columns of Table 5.6 demonstrate that the scoreboard phosphene simulation of clip art images is more illustrative and easier in recognition compared to that of the real high-resolution images due to the simple representation of the object (i.e., abstract representation) without showing unnecessary details. This is clear in the case of the apple and the car.


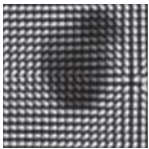

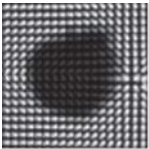
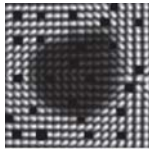



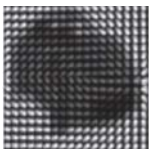
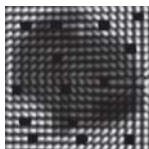
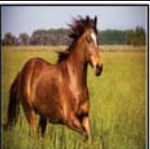


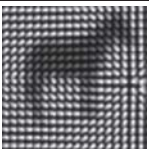
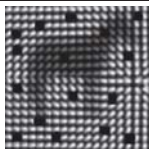

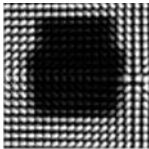

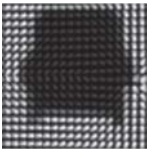
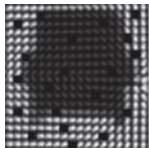
More importantly, objects in the test images that do not belong to the training classes were also successfully converted to their corresponding clip art representation, which can be easily recognized when displayed in phosphene simulation as in the case of the horse and the sofa. In this case, the two sample images took their shape from the most relevant objects that the PVGAN was trained on, where the most similar objects from the training set are the zebra and the chair, respectively. Table 5.6 also demonstrates the impact of the adjustments applied to the clip art images. For the car, it is obvious that the clip art generated by PVGAN has the same orientation (pose) and color as that of the real high-resolution image. Finally, the last column in Table 5.6 shows the phosphene simulation of the clip art images when 20% electrode dropout is applied (i.e., black spots at certain locations in the visual field due to the malfunction of some of the implanted electrodes). However, despite the electrode dropout, the objects are still recognizable.

Table 5.6: Scoreboard phosphene simulation.

	Test Image	Real Image	Generated Clip Art	Clip Art no Dropout	Clip Art with Dropout
New Images from the training classes					
					
Images from new classes not belonging to training classes					
					

In addition to the output of the scoreboard phosphene simulation model demonstrated in Table 5.6, Table 5.7 shows the outcome of using the generated clip art by PVGAN in axon map phosphene simulation. Similar to the observations obtained from Table 5.6, using clip art simplifies the object representation, thereby, makes it easier to recognize the identity of the object. It should be noted here, given the difference in the setup of the two experiments and how the subjects were presented with the images, that none of the subjects was presented with all phosphene simulations appearing on the same row in the table.

Table 5.7: Axon map phosphene simulation.

	Test Image	Real Image	Generated Clip Art	Clip Art no Dropout	Clip Art with Dropout
New Images from the training classes					
					
Images from new classes not belonging to training classes					
					

5.3.5 PVGAN Experimental Results

5.3.5.1 Scoreboard Model Results

To quantify the performance of the proposed approach, we conducted experiments involving four groups of subjects. Figure 5.9 shows the results of the four experiments which comprise displaying scoreboard phosphene simulation of the real images (as a control group), displaying scoreboard phosphene simulation of the real images followed by scoreboard phosphene simulation of the clip art images, displaying scoreboard phosphene simulation of clip art images without dropout and displaying scoreboard phosphene simulation of clip art images with dropout. We examined first the performance of the subjects when presented with images that belong to the same classes used in training PVGAN. First, Figure 5.9a illustrates the recognition time for each group of subjects. The figure demonstrates that the least recognition time is achieved when using clip art representation in comparison to using the real image counterpart (Real Image: 8.99 ± 1.15 sec, Real before clip art: 6.85 ± 0.83 sec, Clip Art no Dropout: 3.05 ± 1.42 sec, Clip Art with Dropout: 3.06 ± 1.12 sec, $P < 1e - 07$, $n = 12$, two-sample t-test). The “Real before Clip Art” group resulted in intermediate recognition time since the real image is first displayed, which spans 5 seconds, before displaying the corresponding clip art image. In addition, no significant difference can be observed between the groups presented with the scoreboard phosphene simulation of clip art without dropout versus those presented with the scoreboard phosphene simulation of clip art with dropout. This indicates that the occurrence of electrode dropout might not hinder the recognition ability of the subjects if clip art representation generated by PVGAN is used.

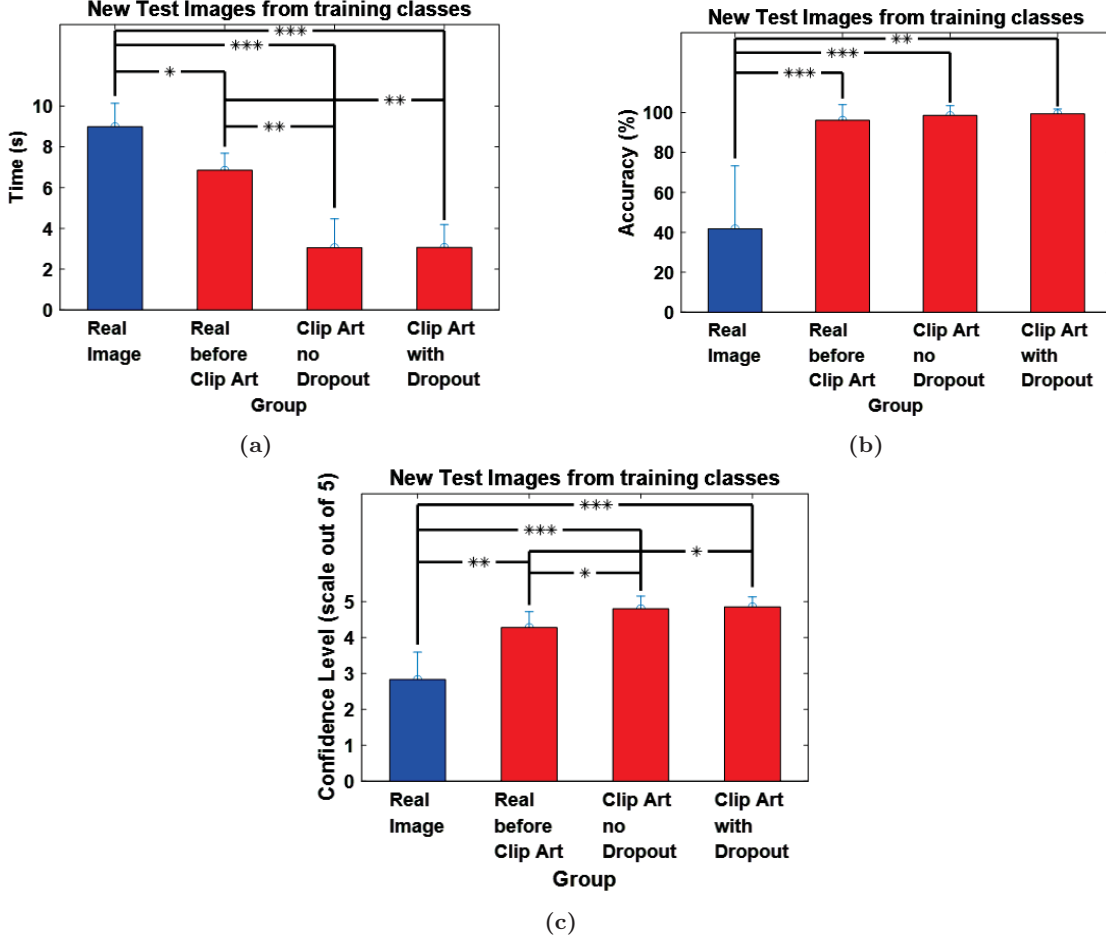


Figure 5.9: Results from scoreboard phosphene simulation for new test images from training classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (*mean ± std*). * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

We also evaluated the ability of the subjects to correctly recognize the presented objects. Figure 5.9b shows the recognition accuracy of each of the four groups of subjects. The figure demonstrates that the experiments including the display of clip art images resulted in similar accuracies, and outperformed that of the scoreboard phosphene simulation of the “Real Image” group (Real Image: $41.76 \pm 31.47\%$, Real before Clip Art: $96.18 \pm 7.81\%$, Clip Art no Dropout: $98.53 \pm 4.93\%$, Clip Art with Dropout: $99.41 \pm 2.43\%$, $P < 1e - 07$, $n = 12$, two-sample t-test). This is consistent with the reduced recognition time observed in Figure 5.9a. We finally quantified the confidence of the subjects in their recognition. Figure 5.9c shows that the confidence level for the groups presented with clip art was significantly higher than that of the “Real Image” group (Real Image: 2.83 ± 0.76 , Real before Clip Art: 4.28 ± 0.44 , Clip Art no Dropout: 4.80 ± 0.36 , Clip Art with Dropout: 4.85 ± 0.28 , $P < 1e - 07$, $n = 12$, two-sample t-test.).

We, then, analyzed test images from new classes that were not encountered during PVGAN training in the four groups of subjects to measure the ability of the model to generalize beyond the training classes. Figure 5.10 shows the results for the four groups using test images from new classes according to the three metrics. The four groups remained with the same properties as that in Figure 5.9. The figure shows that the scoreboard phosphene simulation

of a real-image, as shown in “Real Image” and “Real before Clip Art” groups can be barely identified due to the variety of details in the real-image, whereas much better performance is achieved when clip art is used. This is consistent across the recognition time (Real Image: 9 ± 0.8 sec, Real before Clip Art: 7.4 ± 1 sec, Clip Art no Dropout: 3.8 ± 2.2 sec, Clip Art with Dropout: 4 ± 2.2 sec, $P < 1e - 07$, $n = 12$, two-sample t-test), recognition accuracy (Real Image: $52.9 \pm 30.9\%$, Real before Clip Art: $88.6 \pm 11.8\%$, Clip Art no Dropout: $95.7 \pm 7.9\%$, Clip Art with Dropout: $94.3 \pm 10.2\%$, $P < 1e - 04$, $n = 12$, two-sample t-test), and confidence level (Real Image: 3.2 ± 0.8 , Real before Clip Art: 4.3 ± 0.3 , Clip Art no Dropout: 4.4 ± 0.4 , Clip Art with Dropout: 4.5 ± 0.4 , $P < 1e - 04$, $n = 12$, two-sample t-test).

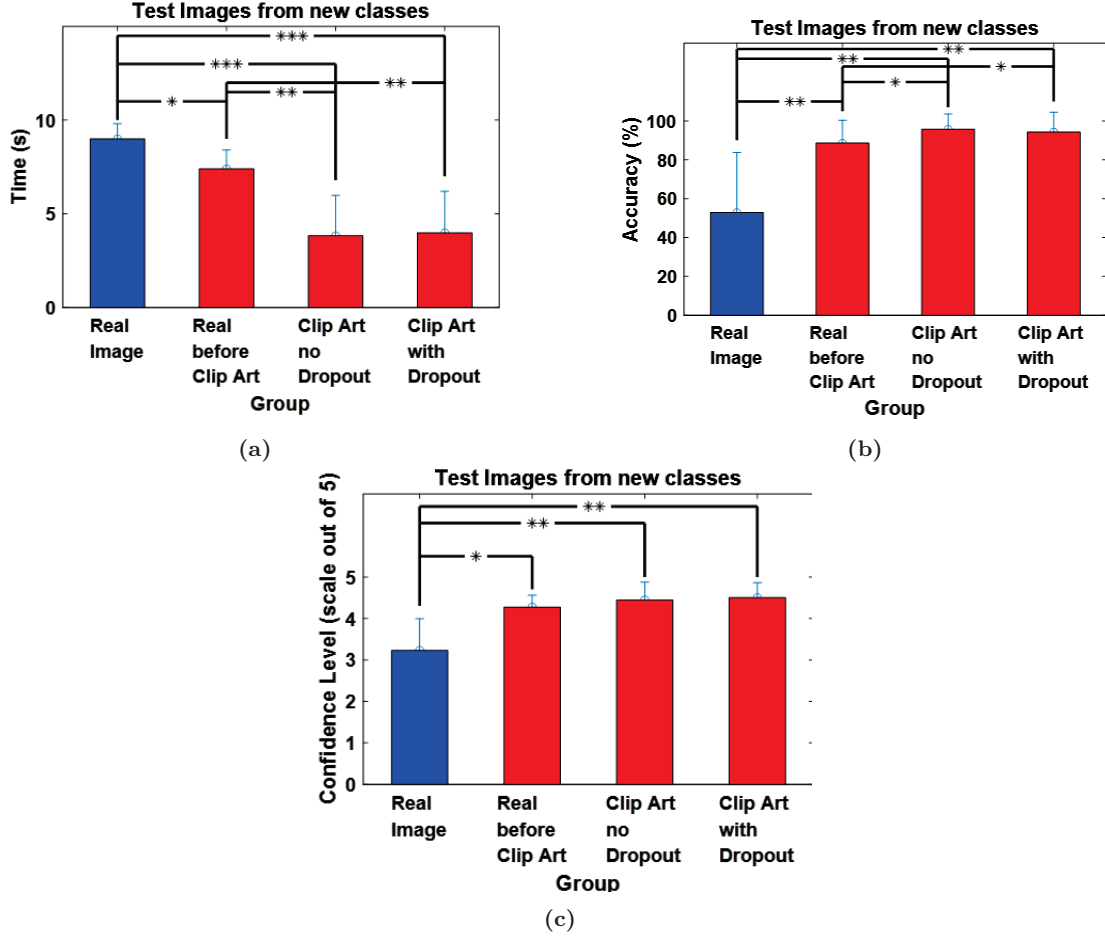


Figure 5.10: Results from scoreboard phosphene simulation for test images from new classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (*mean* \pm *std*). * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

5.3.5.2 Axon Map Model Results

In this experiment, each subject was presented with 24 images comprising the four types of images previously shown in Table 5.7. Figure 5.11 shows the performance averaged across the subjects and types when using the axon map phosphene simulation. Similar to the analysis of the scoreboard model, we examined first the performance of the subjects when presented with images that belong to the same classes used in training PVGAN. Consistent with the results obtained when using the scoreboard model, the least recognition time is achieved when

using clip art representation in comparison to using the real image counterpart (Real Image: 9.63 ± 0.64 sec, Real before clip art: 7.47 ± 0.45 sec, Clip Art no Dropout: 4.37 ± 1.07 sec, Clip Art with Dropout: 4.1 ± 1.05 sec, $P < 1e - 07$, $n = 12$, two-sample t-test). No significant difference can be observed between presenting the subjects with clip art without dropout versus clip art with dropout. Similarly, in terms of the recognition accuracy, Figure 5.11b demonstrates that the approaches of clip art images outperformed that of the direct axon map phosphene simulation of the real images (Real Image: $23.33 \pm 40.41\%$, Real before Clip Art: $100 \pm 0\%$, Clip Art no Dropout: $95 \pm 8.67\%$, Clip Art with Dropout: $98.33 \pm 2.89\%$, $P < 1e - 07$, $n = 12$, two-sample t-test). Finally, Figure 5.11c shows that the confidence level when presented with images including clip arts was significantly higher than that of the real images (Real Image: 1.37 ± 0.06 , Real before Clip Art: 4.4 ± 0.35 , Clip Art no Dropout: 4.73 ± 0.25 , Clip Art with Dropout: 4.53 ± 0.06 , $P < 1e - 07$, $n = 12$, two-sample t-test.).

Testing PVGAN output using test images from new classes that were not encountered during training demonstrated similar enhancement in the performance of the subjects. Consistent results are achieved when comparing Figure 5.12 to Figure 5.11 in terms of the recognition time (Real Image: 9.8 ± 0.35 sec, Real before Clip Art: 7.17 ± 0.06 sec, Clip Art no Dropout: 3.87 ± 0.15 sec, Clip Art with Dropout: 3.7 ± 0.46 sec, $P < 1e - 07$, $n = 12$, two-sample t-test), the recognition accuracy (Real Image: $10 \pm 17.32\%$, Real before Clip Art: $95 \pm 8.66\%$, Clip Art no Dropout: $100 \pm 0\%$, Clip Art with Dropout: $100 \pm 0\%$, $P < 1e - 07$, $n = 12$, two-sample t-test), and the confidence level (Real Image: 2.1 ± 0.98 , Real before Clip Art: 4.67 ± 0.29 , Clip Art no Dropout: 4.27 ± 0.49 , Clip Art with Dropout: 4.43 ± 0.21 , $P < 1e - 07$, $n = 12$, two-sample t-test). These results confirm the utility of using clip art representation in both phosphene simulation models.

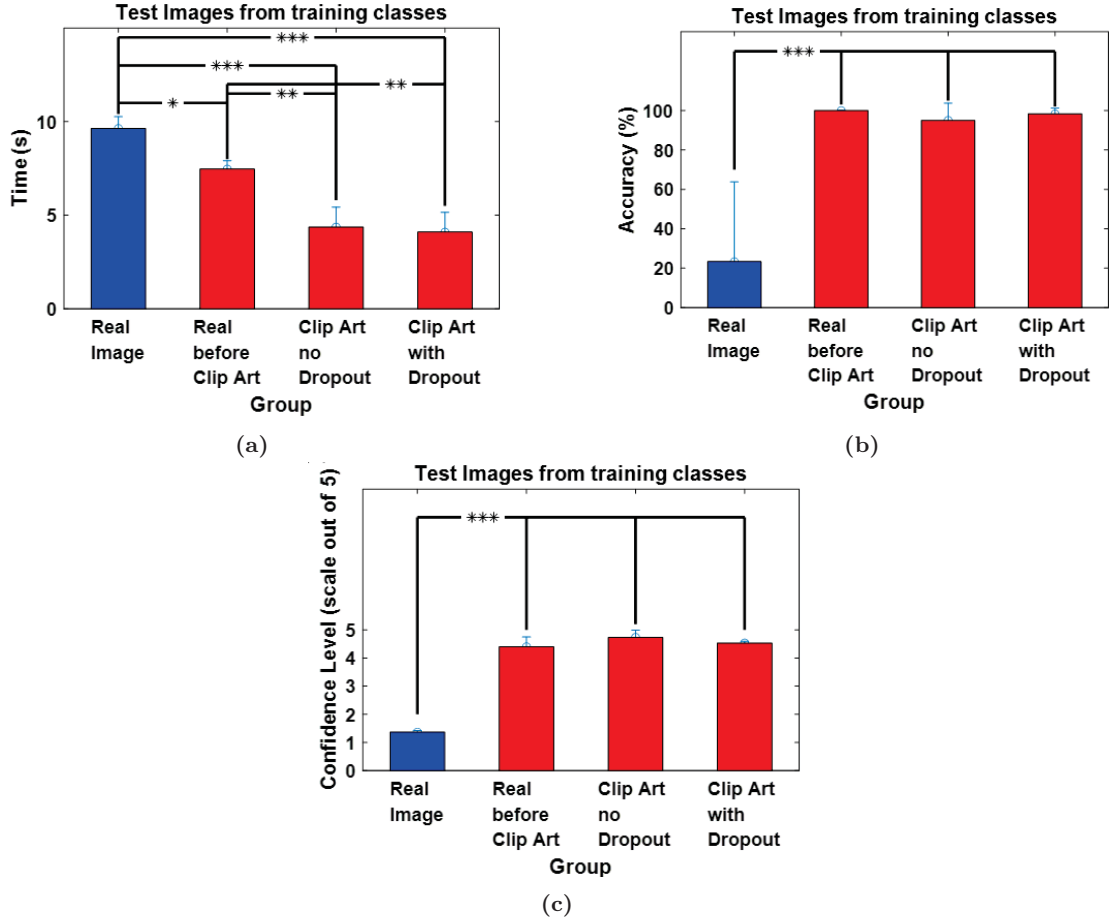


Figure 5.11: Results from axon map phosphene simulation for test images from training classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level ($mean \pm std$). * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

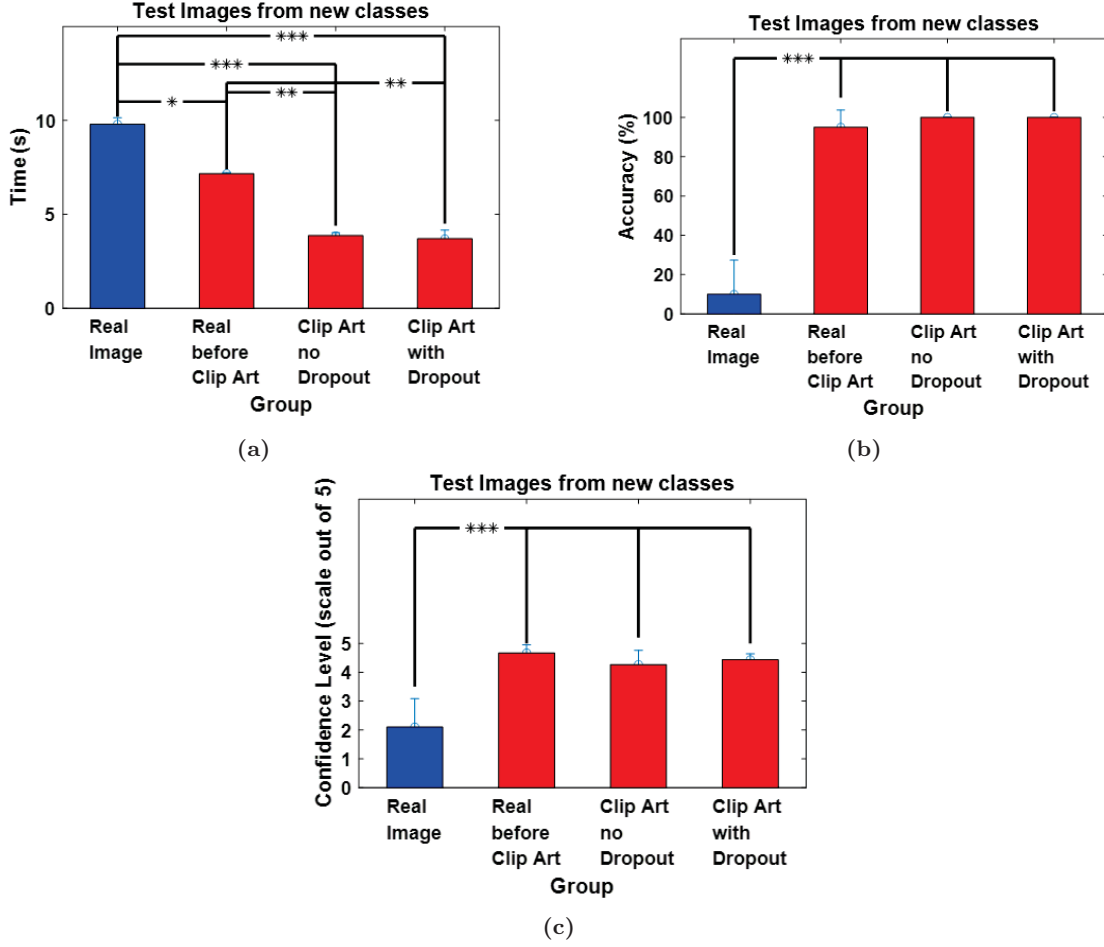


Figure 5.12: Results from axon map phosphene simulation for test images from new classes. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level (*mean ± std*). * $P < 0.05$, ** $P < 1e - 04$, *** $P < 1e - 07$, two-sample t-test.

5.3.6 Performance Analysis

5.3.6.1 ClipArtGAN

We presented ClipArtGAN; a GAN trained to generate clip art images from their corresponding photos. We examined the performance of CycleGAN and Pix2Pix in this task in addition to three different training datasets. The generated ClipArt images from the Pix2Pix GAN using each of the three training datasets show that the model was able in all cases to generate clip art representations of test images that match the identity of the objects. The generated clip art using FC dataset does not result accurate representation of the shape, color or orientation of the test image’s object. However, the FC-OCM dataset resulted in better representation in terms of the color and the orientation, but without maintaining shape matching of the test image’s object. Finally, the HOG-OCM dataset gave the best results as it allowed the Pix2Pix GAN to generate clip art that are visually similar in terms of shape, color and orientation to that of the test image. This is due to the variety of choices that the GAN was trained on when providing the GAN with multiple representations of clip art images for photos within the same class.

The performance of Pix2Pix was significantly better than CycleGAN as quantified by different evaluation experiments, where the average accuracy obtained using CycleGAN is 22%, which is short by 58% compared to that of Pix2Pix. This could be explained given that the Pix2Pix works on pairs of images from different domains where the number of objects in the first domain should be similar to that of the second domain such as mapping one apple fruit to one orange fruit. However, in CycleGAN, the number of objects does not matter; instead, the mapping of the shape of one domain to another domain is done such as mapping an apple to a set of oranges. We are dealing with images that have exactly one object of interest as extracted using a bounding box. In addition, the average accuracy of the generated images from ClipArtGAN is 79% as evaluated by participants, which is more than that of the generated images from Google Images by 37%.

This shows that Google Images was unable in most of the cases to get the corresponding clip art for a certain image in case an image is uploaded to it not searched for its name in the search bar. This could be attributed to the fact that Google Images searches for visually similar images regardless to the selected image type. Finally, the average accuracy of Google Images in recognizing the objects in the generated clip art images generated from CycleGAN using the three datasets is 21% which is less than that of the Pix2Pix GAN by 57%. We attribute the success of ClipArtGAN in generating representative clip art to the use of Pix2Pix in addition to the training datasets. The latter represented one major challenge since we focused on classes that are of different categories to train the model with variety of objects. In addition, we searched for a dataset that includes images that have only one object of interest on a background. This was challenging as the majority of the already available datasets include various objects of interest. Thus, we collected the datasets from various sources to ensure that each image in the dataset includes only one object of interest appearing on a background. Better results could be even achieved if datasets with larger number of images are used to train ClipArtGAN. In addition, utilizing more computational resources such as powerful GPUs could also contribute to the enhancement of the results.

5.3.6.2 PVGAN

We proposed PVGAN as a novel deep learning approach for object simplification that could be used for easier object recognition for visual prostheses users. The model was specifically trained using images of objects that are relevant to visual prostheses users. To prepare the data to train PVGAN, pose and color adjustments were used to match the clip art to the real object. This resulted in generating clip art images for newly seen objects that PVGAN was not trained on, that are correctly oriented based on the orientation of the high-resolution object. This is critical to help visual prostheses users to figure out the direction of moving objects such as, for instance, vehicles in the streets.

Introducing the use of GANs for object simplification represents a novel direction compared to previous efforts that attempted to enhance object representation for visual prostheses users. For instance, background subtraction was proposed to enhance the quality of the perceived image. However, the details of the objects in the image are not feasible to be shown due to the limited number of electrodes that hinders the details from being recognized [60]. Moreover, in [61], segmentation was used to separate foreground objects from the background to allow

an enhanced perception of images. This will enable objects to be segmented by means of a certain property such as region-based segmentation or color-based segmentation. However, those objects are still represented with the full details which will be difficult to determine in the perceived low-resolution images. Another approach proposed by Sanchez-Garcia et al. used grayscale histogram equalization to improve the contrast of the input images [82]. Edge enhancement was also introduced by Dowling et al. to enhance the recognition of objects when displayed in the low-resolution environment which reduces the amount of needed information for recognition [150]. While these approaches represent different enhancement directions, they all rely on representing the full details of the object. On the other hand, in PVGAN, we aim to simplify the object representation for better visualization to match the low-resolution representation perceived by visual prostheses users.

PVGAN utilized Pix2Pix, a well-known GAN architecture for image-to-image translation. While other models exist in the literature that could be used in the same task, Pix2Pix seemed the most appropriate given that it is relatively simple and capable of generating large high-quality images across a variety of image translation tasks.

To analyze the phosphene simulation of the different types of images (i.e., real high-resolution images and clip art images), scoreboard phosphene simulation and axon map phosphene simulation models were used. These models represent different types of phosphenes that have been reported by implanted patients. Therefore, evaluating PVGAN using both simulation models would represent a rigorous assessment regardless of the phosphenes shape. In both models, the phosphene simulation of the real high-resolution images resulted in ambiguous and undefined objects. This represents the current perception available in visual prosthetic devices, representing a limitation of current visual prostheses. This could be attributed to the limited number of electrodes used to represent the visual field. The performance of the subjects using both simulation models confirmed our hypothesis that using clip art representation is better in representing the objects regardless of the phosphene simulation model used. Subjects were able to recognize the objects in a fast manner and their confidence was extremely high in identifying the identity of an object. This is due to the fact that it is better and easier to see an image that has only a simple representation of the object of interest as opposed to seeing a real image that is full of irrelevant details. Moreover, even in the case of having electrode dropout, understanding and recognizing a clip art image was not significantly affected by the presence of dropouts, indicative of the efficacy of the proposed approach. This could enable visual prostheses users to better interpret the viewed scenes, giving them more independence and confidence.

5.3.7 Conclusion

Visual prostheses have demonstrated significant success in recent years in compensating for the loss of vision in blind patients. However, multiple challenges have been observed that need to be addressed for better adoption of this technology. While some of these challenges could be addressed through the development of the implant hardware, image processing could still be utilized to provide better perception using the current technology. We proposed a clip art generation deep learning model that used GANs; termed ClipArtGAN. This model can be utilized in any aspect in addition to visual prostheses to enable clip art images generation.

The model comprises a Pix2Pix GAN that is trained with a color and orientation adjusted images based on HOG features. The proposed model was evaluated using three different experiments. In all experiments, ClipArtGAN demonstrated its superiority in comparison to changing the GAN model to be CycleGAN and changing the training dataset, and in comparison to Google Images recognition. To our knowledge, this is the first model that is capable of automatically generating clip art images for any arbitrary shape. We next modified the proposed ClipArtGAN model to be used to improve object recognition for visual prostheses users; named PVGAN. PVGAN is similar to ClipArtGAN in terms of the architecture, however, the trained classes are more diverse in PVGAN compared to that of the ClipArtGAN. Another two sets of experiments were conducted to measure the performance of PVGAN, the first set of experiments utilized scoreboard phosphene simulation, while the second experiment utilized axon map phosphene simulation. All the experiments were conducted with performance measured through three evaluation metrics (time to decision, recognition accuracy and confidence level). The introduction of clip art in the phosphene simulation, either scoreboard or axon map, showed easier, faster and more accurate recognition compared to that of real high-resolution images. The results obtained showed that the generated clip art images used in the phosphenes simulation gave outstanding results in objects' recognition both with and without dropout added to the phosphene simulation.

6

Mixed Reality Real-Time Simulation

6.1 Introduction

Given the poor spatial and radiometric resolutions of current visual prostheses system, that gets more problematic with the existence of electrode dropouts, we propose a system for enhancing object localization and object recognition. Object recognition refers to the ability of the users to correctly identify an object's identity, whereas object localization refers to the ability of the users to correctly identify the location of an object and grasp it. The first enhancement is the usage of clip art representation in place of the actual object as a scene simplification method to allow better recognition as demonstrated in Chapter 5. Second, we utilize an edge enhancement technique along with corners enhancement to sharpen the edges of the object for better detail preservation. While edge enhancement 7 techniques were examined before in the context of enhancing prosthetic vision [150], they were not 8 combined with using clip art representation. Finally, we apply the dropout handling technique introduced in Chapter 4 to support preserving the maximum possible numbers of phosphenes despite the existence of the malfunctioned electrodes that caused dropout [151]. The proposed approach is examined in a real-time mixed-reality (MR) environment that simulates perceived vision by visual prostheses users.

6.2 Methods

6.2.1 System Overview

The proposed enhancement techniques were examined using simulation of prosthetic vision presented through a mixed reality (MR) setup as shown in Figure 6.1a. The figure shows one sample object (a cup) that was presented to the subject. The camera in the setup takes a picture of the environment, where only 20° of the visual field is displayed inside the VR headset. The real object is displayed in its phosphene simulation allowing the participant to move freely to locate the object in its actual location. Phosphene simulation of the enhanced version of the object, including translating the enhanced image within the presented view to

minimize the number of dropouts, is displayed to allow better recognition ability. The dropout locations were randomized across subjects. However, for each subject, the dropout locations remained fixed throughout the experiment to mimic the malfunctioning of specific electrodes in an implant.

The captured image is pre-processed through multiple stages as shown in Figure 6.1b. First, an image is captured using the mobile camera that acts as the PC webcam. The image is then converted from RGB to grayscale to mimic the colors used by visual prosthetic users. Next, to simulate the visual field that visual prosthetic users encounter, a circular mask is applied. The radius of the mask reflects the 20° of the visual field given that the visual field in a prosthetic vision device is approximately 20° from the complete visual field [152]. Element-wise multiplication is performed between the mask and the pre-processed image so that only the part of the image residing inside the visual field is displayed. Furthermore, dropouts are added at random locations to mimic the malfunctioning of the electrodes at random positions.

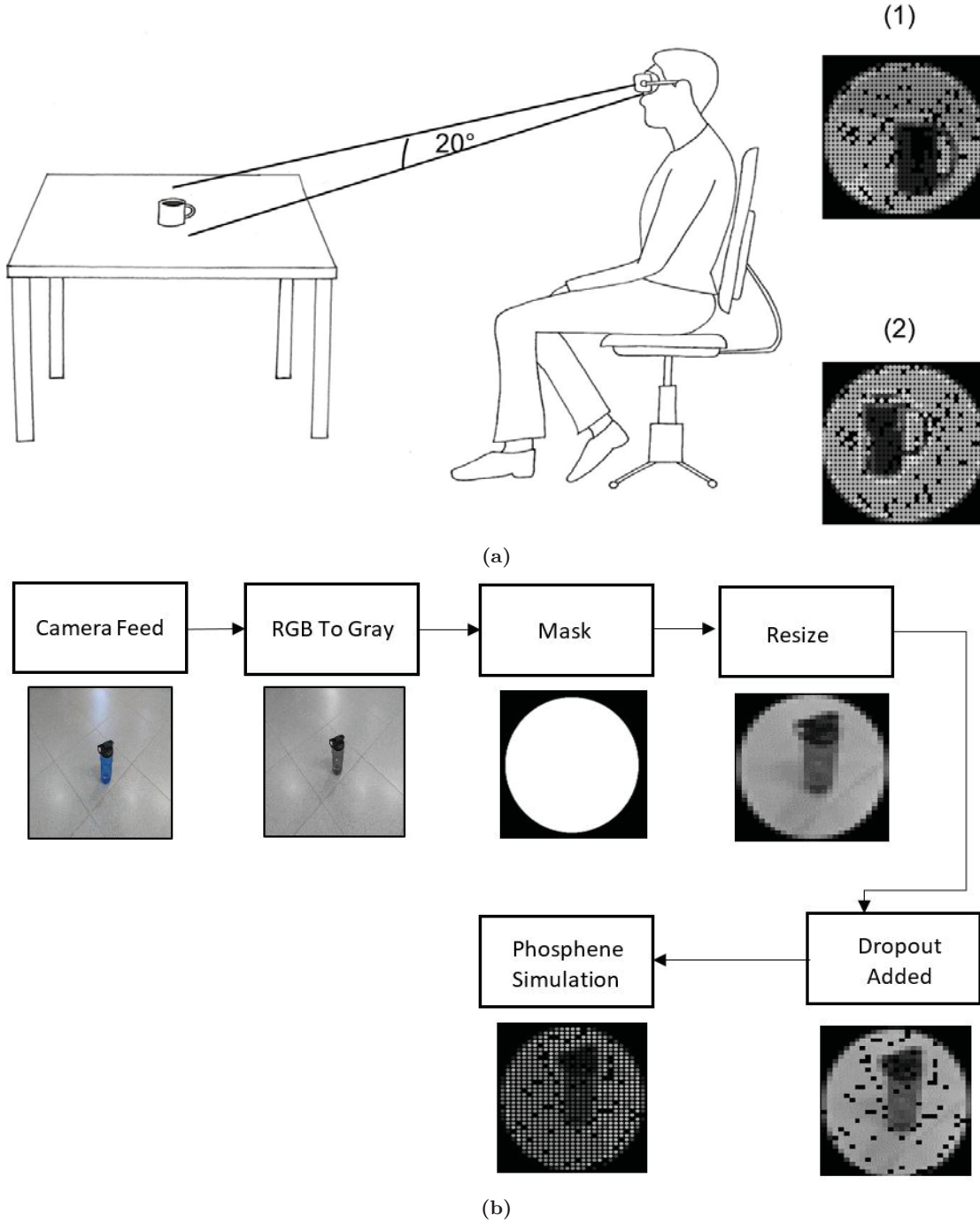


Figure 6.1: Displayed Phosphene Simulation. (a) Sample Phosphene Simulation Perception in Experiments (1) Before Image Enhancement. (2) After Image Enhancement. (b) Visual Field Adjustment in Phosphene Simulation.

6.2.2 Phosphene Simulation

The phosphene shape used was a round shape that matches the common form for phosphenes simulating the actual look of the phosphenes without any change in the stimulation amplitude and with ideal current set [153]. A squared grid was used in the phosphene simulation since it simulates the common grid used in visual prostheses [90]. Furthermore, the distance between

each two successive phosphenes was set to 0.5 [90]. A dropout rate of 10% was used, where the dropout phosphene color was set to a black color [83]. We used 8 (3 bits) gray levels since the number of successfully distinguishable gray levels by real patients is in the range of 4 to 12 levels [92]. To map the initial 256 (8-bits) gray values to their corresponding gray value, a mapping scheme of 0, 36, 72, 108, 144, 180, 216 and 252 was utilized [96].

6.2.3 Proposed Enhancement Techniques

6.2.3.1 Clip Art Representation

To enhance object recognition, we propose the utilization of a simplified version of an image, which is the clip art, to enable abstract representation of the image given the low spatial and radiometric resolutions [154]. The clip arts of the utilized objects in all of the experiments were selected, where the best shaped clip arts that easily identify an object are collected. The clip art size is adjusted to match that of the real object size. The clip art representation was mainly utilized to enhance the ability of the user to recognize the objects.

6.2.3.2 Edge Enhancement

To enhance object localization, we use edge detection to emphasize the borders of the object of interest to facilitate its recognition. Canny edge detection was used to identify the edges in the object of interest where the edges in the vertical and horizontal directions are detected [155, 156]. The derivative of a Gaussian filter is used by the edge to determine the gradient. This technique detects both strong and weak edges using two thresholds and includes weak edges in the output if they are connected to strong edges. The Canny approach uses two thresholds, which makes it less susceptible to noise than the other methods and more likely to identify real weak edges. This thresholding is performed to the thinned edge magnitude image utilizing two edge strength thresholds named hysteresis. All pixels are candidates to be edge pixels where an edge pixel is a pixel that is above the low threshold of value 0.1 which can be connected to any arbitrary pixel above the high threshold of value 0.15 through a chain of edge pixels [156]. Moreover, non-maximal suppression is applied to thin the edges in which surviving candidate pixels are determined. Finally, the original grayscale image was added to the edge version of the object so that the edges in the object of interest will be sharpened.

6.2.3.3 Corner Enhancement

For corner detection, we used FAST for key points' feature extraction, where a minimum accepted quality of corners of 0.1 and a minimum intensity of 0.2 were used [157]. FAST utilizes a circle comprising 16 pixels to determine corners from the candidate points. Every pixel is labelled from 1 to 16 clockwise. A corner pixel is a pixel where its intensity plus a threshold value t , is darker than a set of N pixels in the circle or its intensity minus a threshold value t of 0.2, is brighter than a set of N pixels. We utilized a value of N of 12 since that it is the most commonly used value so that the number of detected corners will be reasonable [158]. We utilized FAST algorithm to develop an interest point detector for utilization in real time mixed reality simulation.

6.2.4 Tools Used

To create a real-time mixed reality (MR) setup, the following was performed: The Trinus VR application was used to display the phosphene simulation on a mobile screen that is placed inside the VR set (Electro Shinecon VR Box 3D headset, Dongguan Shinecon Industrial Co., Ltd, China). To allow the mobile camera to act as a PC webcam, a mobile application (IP Webcam) was used to enable capturing real-time images, sending them to the PC wirelessly to prepare the phosphene simulation and then, displaying the phosphene simulation on the mobile screen. The number of phosphenes used was 32×32 which is the threshold of scene recognition [44]. Moreover, a visual field of 20° was used to mimic the legal blindness threshold [97]. The process for each phosphene simulation takes 0.8 sec, which is considered fast given the larger number of pixels used compared to that of a typical visual prosthetic device. The Trinus VR displayed the mask in the phosphene simulation as an ellipsoidal-like one, not circular. Therefore, to solve this issue, we manipulated the generated phosphenes to be circular by creating an ellipsoidal-like mask in the phosphene simulation so that a circular mask will be displayed using Trinus VR in the VR headset.

6.2.5 Experimental Setup

Twelve experiments were conducted on corrected vision/normally-sighted subjects. Each of the twelve experiments comprised five participants giving a total of 60 participants involved in all of the experiments with a range of age from 19-36 years old (23.37 ± 4.1 years old) for both genders (34 males and 26 females). All the subjects signed a consent form indicating their confirmation to participate voluntarily in the experiments. The answers of the subjects were recorded through an audio recorder to avoid any human error in miscalculating the actual time for recognizing or localizing a certain object. Each of the participants was asked about their visual acuity and the reported visual acuity was recorded. However, visual acuity would have no impact on the results given the very low-resolution used in the phosphene simulations to mimic prosthetic vision resolution. Different subjects were involved in each experiment to avoid any learning effect that might arise due to prior knowledge of the displayed scene from a previous experiment. A demo experiment was presented to the participants before all the actual experiments to introduce the subjects to the phosphene simulation interpretation to be able to perceive the images used in the actual experiments in an easy manner.

The experimental paradigm used in all of the conducted experiments is shown in Figure 6.2a, where the subject first wears the headset, then there are two phases that the subjects pass through. In Phase 1, a black screen is displayed for 10 sec to mimic the real environment that a real patient encounters before the visual prosthetic device is turned on and before being presented with any objects. Then, the phosphene simulation of the scene is presented to the subject before placing any object to explore the environment in front of the subject. In Phase 2, a black screen is displayed again for 10 secs, but this time it acts as a cue that a new object is currently being placed in the scene. The subject is then allowed a maximum duration 120 sec for the single object recognition experiment and localization, 240 sec for the multiple object recognition and localization experiment, and 180 sec for the navigation experiment. Phase 2 keeps repeating again as long as a new object is being placed in the scene. Three setups were used in the experiments. The first setup, shown in Figure 6.2b, is for single object

recognition and localization where the participant is seated with the object being placed on a table with a white background to enhance the contrast between the background and the foreground. This setup comprised 8 experiments including an experiment on a control group. The second setup, shown in Figure 6.2c, was for multiple objects recognition and localization. This setup comprised 2 experiments including an experiment on a control group. In the first two setups, the subjects were seated on a chair with the freedom of upper body movement such as getting closer to the object to allow zooming in the object or moving away from the object to allow zooming out of the object. The third setup, shown in Figure 6.2d, was for objects recognition and localization while navigating in an indoor environment. This setup comprised 2 experiments including an experiment on a control group.

In all experimental setups, all participants were asked to name the object before attempting to grasp it. This is to ensure that the correct object recognition is only due to the ability of the participant to visually recognize the object, and not due to feeling the object when touching it during the grasp attempt. A set of questions was asked to the participants in all of the experiments as follows:

- Are you able to detect any object? a) yes or b) no
- Describe the geometry of the shape you see: a) rectangular, b) curve-like or c) other description
- What is the shape?
- Can you determine its location? a) yes or b) no
- How confident are you with your answer on a scale from 1-5, where 1 is the least confidence, and 5 is the highest confidence?

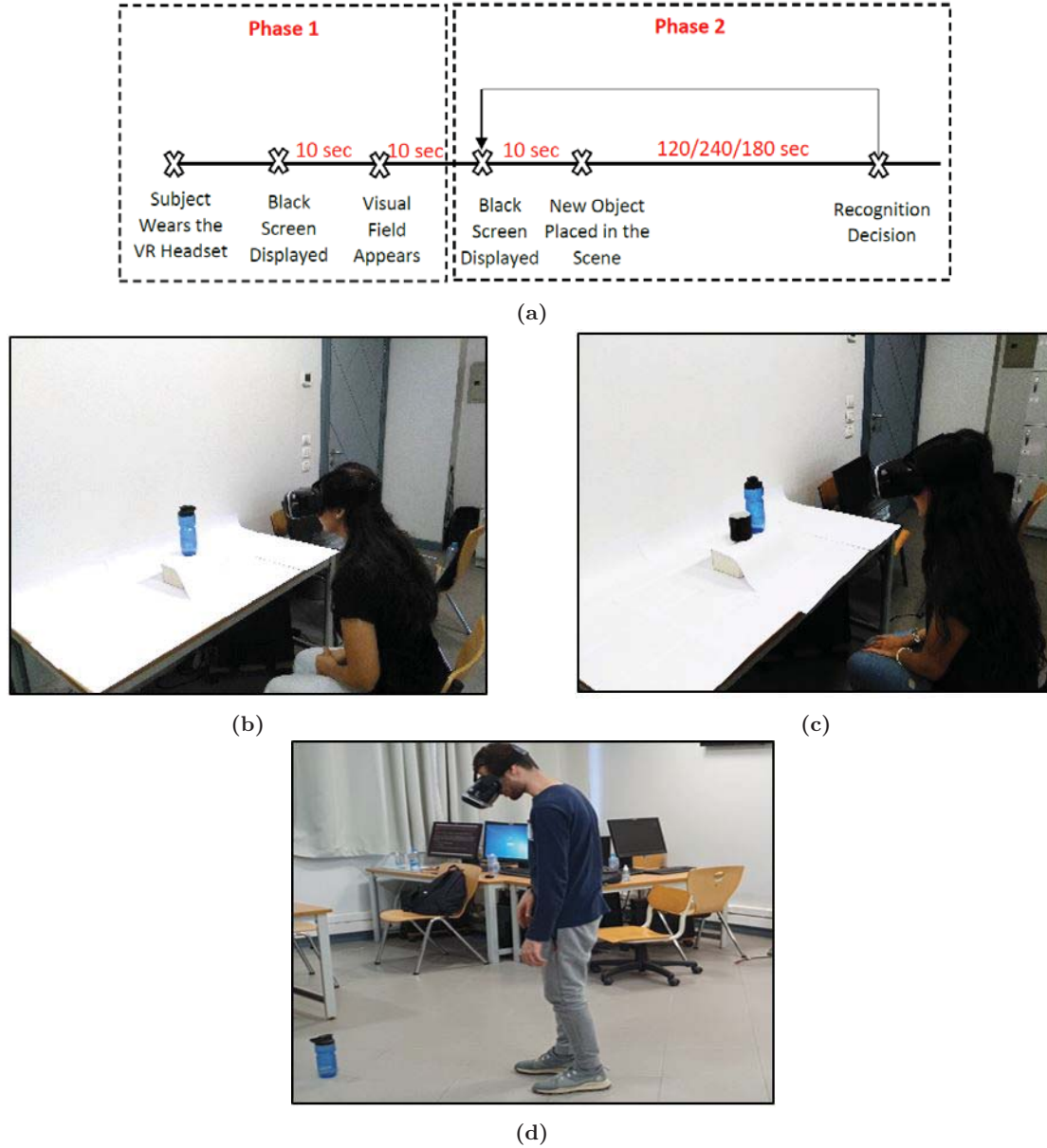


Figure 6.2: Experiments Set-Up. (a) Experimental Paradigm Timeline. (b) Single object Experiment Setup. (c) Multiple objects Experiment Setup. (d) Navigation Experiment Setup.

6.2.5.1 Single Object Recognition and Localization Experiments

Five objects were used, which are cup, bottle, banana, car toy and laptop, in all of the eight experiments. The duration given for the participants for both object recognition and localization per one object was 120 sec (i.e., 2 minutes). The participants wore the VR Box headset and were asked the aforementioned set of questions during the course of the experiment. The first experiment was the control group experiment which involved 2 males and 3 females. This group was presented with the phosphene simulation of the real object placed in front of them, with randomly generated dropouts added once at the beginning of the experiment for each subject and without using any enhancement technique. The second experiment was conducted on 4 males and 1 female, where the dropout handling technique was performed.

This was in addition to the clip art representation of the real object, where the clip art object was resized and translated to the optimal location in the visual field. The third experiment involved 3 males and 2 females in which dropout handling was performed in addition to edge enhancement, applied to the real object image, using Canny edge detection technique. For the fourth experiment, it involved 2 males and 3 females, where dropout handling was performed in addition to FAST corners detection to preserve the important details in the object. The fifth experiment comprised 4 males and 1 female where edge enhancement and FAST were applied to the real object without dropout handling. In the sixth experiment, 3 males and 2 females were involved in which clip art and edge enhancement were performed without dropout handling where the clip art is resized and translated to the location of the actual object in the visual field to maintain the actual location of the real object. The seventh experiment involved 4 males and 1 female where clip art and FAST were performed without dropout handling. Finally, the last experiment comprised 3 males and 2 females in which all enhancement techniques including clip art, edge enhancement, FAST and dropout handling are applied to the image. All the images displayed in all the experimental setups were represented in the phosphene simulation.

6.2.5.2 Multiple Objects Recognition and Localization Experiments

Three objects were taken in pairs at a time to be used in this type of experiments which are a bottle and a cup, a bottle and a banana, and a banana and a cup. The first experiment, involving 3 males and 2 females, was the control group experiment representing the phosphene simulation of the actual scene. The duration given for the participants for both object recognition and localization per one pair of objects was 240 sec to match that of the duration used in the “All Enhancement Techniques” experiment. The second experiment, involving 2 males and 3 females, used all the enhancement techniques discussed before. Since the participant has the freedom to move his/her head to capture any of the two objects that are displayed at a time, You Only Look Once (YOLO) deep learning model is used to detect the object and get the corresponding clip art [65]. In the single object experiments, only one object is displayed at a time and the order of displaying the objects is fixed across all participants, so the corresponding clip arts were easily determined without the need of waiting for YOLO to detect an already known object. Similar to the control group experiment, a duration of 240 sec was used to be able to both recognize and localize each object per one pair of objects. The 240 sec are utilized to give more chance for the participant to keep moving back and forth and left and right until being able to locate the full shape of an object so that YOLO is able to detect the object correctly and, therefore, the corresponding clip art will be retrieved. Once an object is detected by YOLO, the corresponding clip art is displayed in phosphene simulation. The best clip art shapes for the 80 classes used in the YOLO model were pre-determined, where each image was named based on the identity of the object. Then, the name (i.e., label) of the detected object is taken and compared to the prepared labels to find the corresponding clip art of the query real object. The clip art replaces the actual objects in the scene after applying edge enhancement, FAST and dropout handling. The dropout handling in this case was performed for each of the two bounding boxes of the detected objects separately. Objects were then translated one at a time. Each bounding box is translated to the ideal location

(i.e., the location that fully occupies the object) in the visual field that contains the minimum number of dropouts.

6.2.5.3 Object Recognition and Localization during Navigation Experiments

Three objects were used in these experiments which are a backpack, a bottle and a chair. Since during navigation, visual prostheses users might encounter objects of different sizes, so we used an object of a big size (the chair), an object of a medium size (the backpack) and an object of a small size (the bottle). The duration given for the participants for both object recognition and localization per one object was 180 sec (i.e., 3 minutes). Thus, one extra minute was added for each object to give the participant the chance to walk to the object location (i.e., 2 minutes + 1 minute = 3 minutes per object) compared to that of the single object experiments where no navigation was needed. This unifies the time across all other experimental setups. The first experiment, involving 3 males and 2 females, was the control group experiment without any enhancement applied to the actual image that is displayed in phosphene simulation. The second experiment, involving 2 males and 3 females, was the experiment that applied all the enhancement techniques. In this experiment, the clip art of the displayed object appears in phosphene simulation each 10 sec until a new object is placed, to give the chance to the subject to navigate and see the actual object in phosphene simulation. In both experiments, the subjects were asked to, first, move through the environment while looking at the floor, before any of the three objects is placed on the floor, to give them the chance to interpret how the floor would look like via phosphene simulation. This is to help them know that an object has been added by contrasting the difference in the perceived scene.

6.2.6 Evaluation Metrics

Five evaluation metrics were used in this study to evaluate the performance of the subjects in the experiments, where three of them were used for object recognition evaluation, while the other two were used for object localization evaluation. The evaluation metrics that were used in the object recognition are the recognition time measured in seconds, the recognition accuracy measured as the percentage of correctly recognized objects out of the total number of objects, and the confidence level on a scale from 1 to 5. The evaluation metrics that were used to measure the performance of the participants in objects localization were the grasping accuracy denoting the percentage of the correctly grasped objects with respect to the total number of objects, and the grasping attempt time measured in seconds. For the grasping attempt time, it denotes the time at which the subject was extremely near to the object (i.e., within 10 cm from the object) or was able to grasp the object correctly without attempting to touch the object first to localize it. The 10 cm margin of error was utilized as a threshold for subjects who were relatively close to grasping the object, before mistakenly hitting the object leading to its fall. It should be noted that for the grasping accuracy, it was considered to be 100% if the object was correctly grasped. In case an object was not correctly grasped but the subject was 10 cm away from the object in their grasping attempt time, they were given a grasping accuracy of 0%.

6.3 Results

6.3.1 Enhancement Techniques and Phosphenes Simulation Outcome

To illustrate the shape of the image after applying edge sharpening, FAST and clip art representation, Figure 6.3a shows the input image after applying each of the three methods. It can be observed that the edges are sharpened surrounding the bottle making it easier in recognition due to the outline given to the bottle that resulted from the edge enhancement. Applying FAST emphasized the corners at the bottle's neck which could enable easier identification of the bottle. Finally, using the clip art representation simplifies the bottle further, which in combination with other enhancement techniques is expected to enhance the ability of visual prostheses users to recognize the presented objects.

We next demonstrate the difference between the outcomes obtained from the single objects experimental setups as shown in Figure 6.3b, where the outcomes are shown for different participants. As it is shown, the result from the "Control Group" experiment illustrates what the control group participants see where the real object (i.e., bottle) is barely recognized due to the dropout existence that overlaps with the bottle's body. The utilization of dropout handling and clip art can be shown in the "Dropout Handling & Clip Art" experiment, where the real bottle is replaced by its clip art and translated to the best location within the visual field that has the minimum dropped out locations. Moreover, it can be observed that the addition of the dropout handling and edge enhancement to the real object, as shown in "Dropout Handling & Edge Enhancement" experiment, demonstrates that the edges of the bottle are sharpened and the real bottle is translated to the best location within the visual field with minimum dropouts. Furthermore, applying dropout handling and FAST, as demonstrated in the "Dropout Handling & FAST" experiment, highlights corners at the neck of the bottle in addition to the translation of the real bottle to the location with minimum dropout. In addition, further enhancement to the real bottle can be perceived when utilizing edge enhancement and FAST to the bottle, as indicated in the "Edge Enhancement & FAST" experiment, where the edges of the bottle along with its corners are sharpened. This enhancement was further improved when using the clip art of the bottle along with edge enhancement, as revealed in the "Clip Art & Edge Enhancement" experiment, where the clip art of the bottle along with the edges of the clip art are sharpened as shown. Similarly, the usage of clip art along with FAST, as shown in the "Clip Art & FAST" experiment, sharpens essential corners pixels in the bottle. Finally, the utilization of all the aforementioned enhancement techniques, as illustrated in the "All Enhancement Techniques" experiment, shows the best possible look for any arbitrary object enhancing objects recognition. This was shown when the real bottle was replaced by its clip art representation in addition to sharpening both the edges and the corners of the clip art bottle along with translating the bottle to the place with minimum dropout. The figure demonstrates that the usage of the clip art of the bottle gave better visualization of the bottle where the real bottle was confusion in recognition.

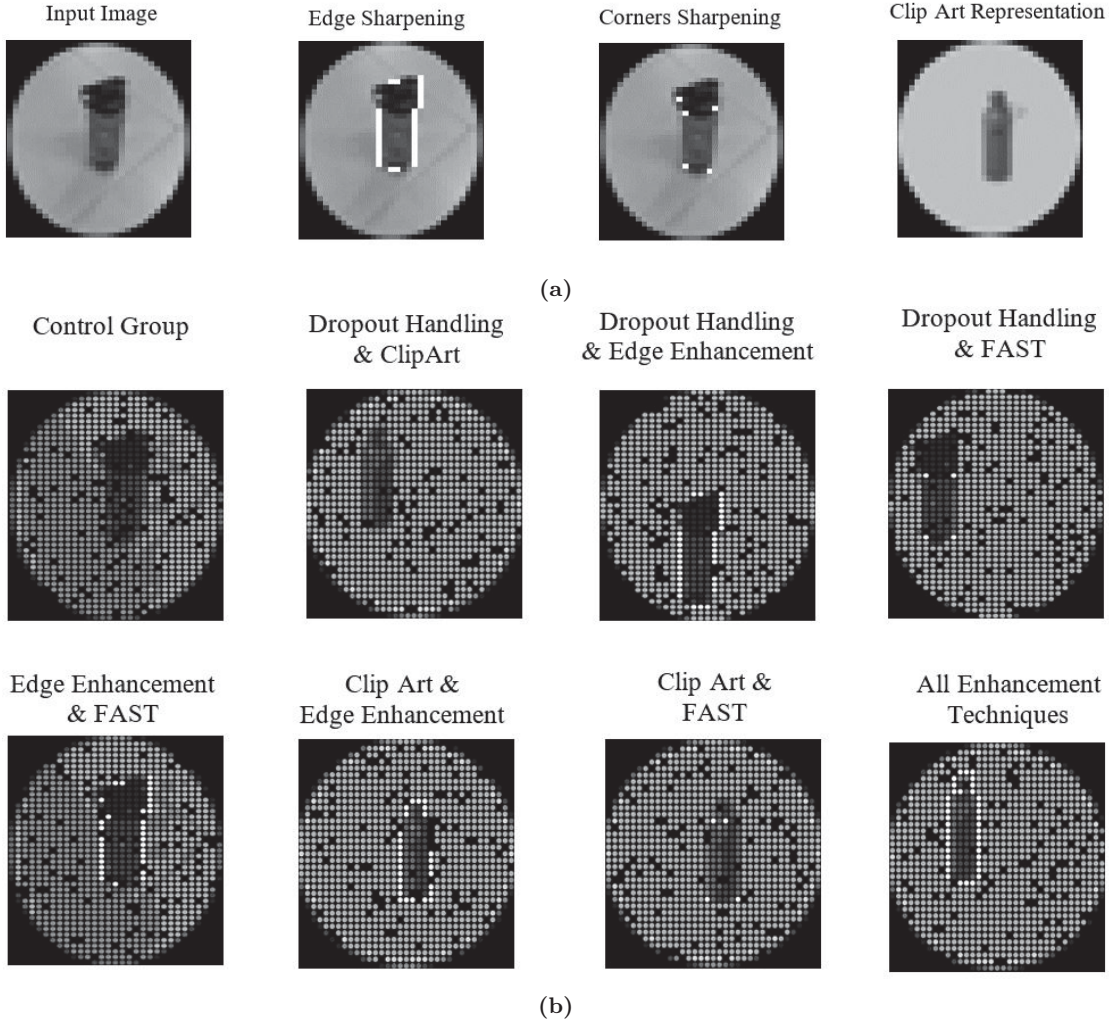


Figure 6.3: Single Objects Experiments. (a) Enhancement Techniques. (b) Phosphene Simulation of the 8 experiments.

To demonstrate the outcomes from the multiple objects grouping experiments, Figure 6.4 shows the outcomes from the two experiments conducted in the multiple objects setup, obtained from two different participants. The “Control group” experiment contains the real bottle and cup without any enhancement technique applied. However, the “All Enhancement Techniques” experiment shows the phosphene simulation of the bottle and the cup after replacing each of the two objects by their corresponding clip art along with sharpening both the edges and the corners in addition to translating the objects to the location with minimum dropout. Similarly, Figure 6.5 shows the outcomes from the two experiments utilized in the navigation system setup for two different participants. The “Control Group” experiment shows the real bottle in the phosphene simulation without the application of any enhancement technique, whereas the “All Enhancement Techniques” experiment shows the phosphene simulation after sharpening the edges and the corners of the clip art object translated to the location with minimum dropouts.



Figure 6.4: Multiple Objects Experiments. (a) Control Group. (b) All Enhancement Techniques.



Figure 6.5: Navigation Experiments. (a) Control Group. (b) All Enhancement Techniques.

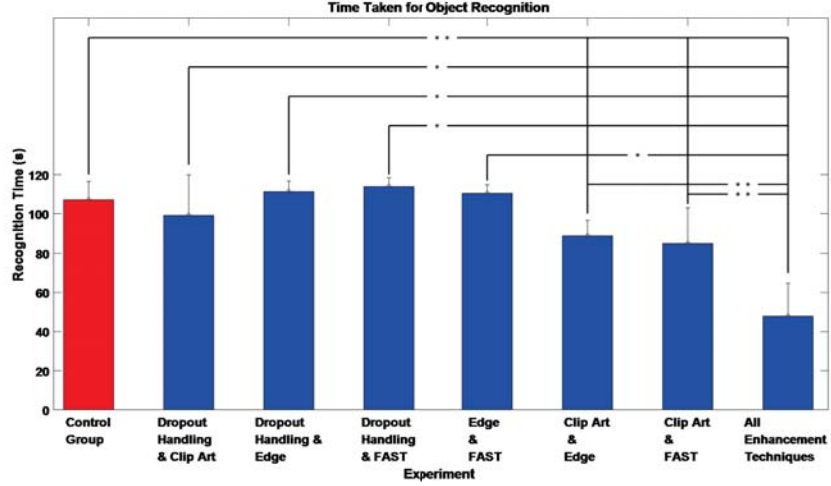
6.3.2 Experimental Results

6.3.2.1 Single Object Experiments

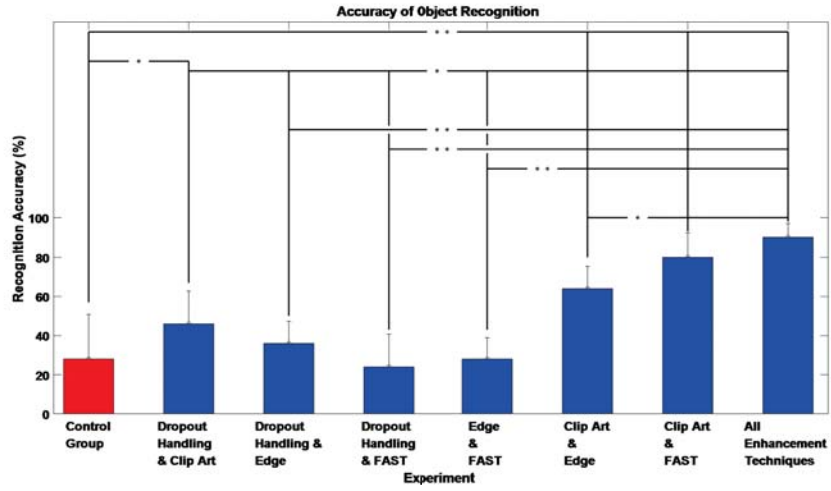
We first examined the performance of the proposed enhancements when presenting the subjects with single objects. Figure 6.6 and Figure 6.8 show the results obtained from the single object experiments (i.e., the 8 experiments). The 8 experiments were given the following names: control group, dropout handling & clip art, dropout handling & edge enhancement, dropout handling & FAST, edge enhancement & FAST, clip art & edge enhancement, clip art & FAST and, finally, all enhancements techniques. Figure 6.6a demonstrates the time taken to correctly recognize the objects. The figure shows that the experiment with all enhancement techniques gives the least recognition time compared to the other experiments (Control Group: 107.08 ± 9.38 sec, Dropout Handling & Clip Art: 99.12 ± 20.73 sec, Dropout Handling & Edge Enhancement: 111.24 ± 5.43 sec, Dropout Handling & FAST: 113.88 ± 4.58 sec, Edge Enhancement & FAST: 110.4 ± 4.38 sec, Clip Art & Edge Enhancement: 88.68 ± 7.88 sec, Clip

Art & FAST: 85.04 ± 17.99 sec, and All Enhancement Techniques: 47.88 ± 16.84 sec). Figure 6.6b demonstrates the accuracy of correctly recognized objects. The figure illustrates that the highest accuracy was also achieved in the experiment with all enhancement techniques compared to the other experiments (Control Group: $28 \pm 22.8\%$, Dropout Handling & Clip Art: $46 \pm 16.73\%$, Dropout Handling & Edge Enhancement: $36 \pm 11.4\%$, Dropout Handling & FAST: $24 \pm 16.73\%$, Edge Enhancement & FAST: $28 \pm 10.95\%$, Clip Art & Edge Enhancement: $64 \pm 11.4\%$, Clip Art & FAST: $80 \pm 12.25\%$, and All Enhancement Techniques: $90 \pm 7.07\%$). Finally, Figure 6.6c demonstrates the confidence level of the participants denoting how confident they were when they recognized the object. Consistent with the recognition time and accuracy results, the figure shows that the experiment with all enhancement techniques gave the highest confidence level compared to the other experiments (Control Group: 1.76 ± 0.68 , Dropout Handling & Clip Art: 2.4 ± 0.75 , Dropout Handling & Edge Enhancement: 2.16 ± 0.22 , Dropout Handling & FAST: 2 ± 0.51 , Edge Enhancement & FAST: 2.16 ± 0.43 , Clip Art & Edge Enhancement: 3.52 ± 0.48 , Clip Art & FAST: 3.8 ± 0.37 , and All Enhancement Techniques: 4.52 ± 0.11). To assess the significance of the results, two-sample t-test was performed across all trials.

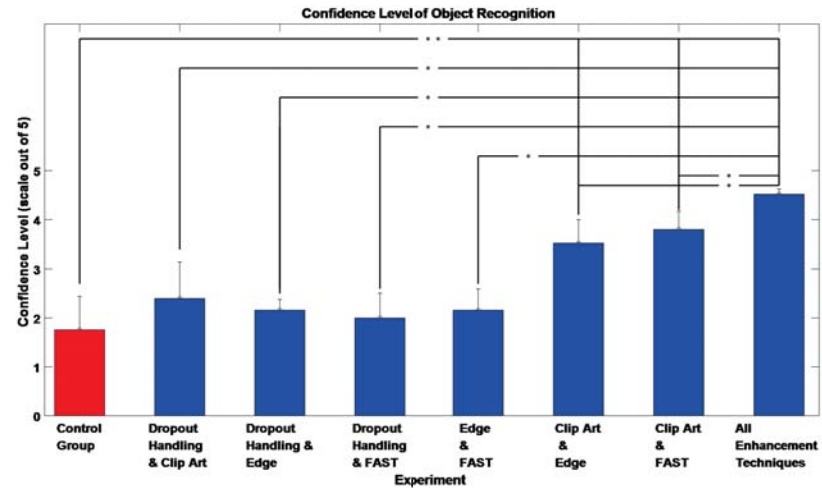
To further assess the ability of different approaches to enhance object recognition, we examined each of the three evaluation metrics (the time taken to recognize the object, the recognition accuracy and the confidence level) for each of the displayed objects. Figure 6.7 indicates that the Banana was the easiest object to recognize, especially when clip art representation was used, achieving the best performance when all enhancement techniques were applied (Recognition Accuracy: 100%, Average Confidence: 4.6). This could be attributed to the distinct curved shape the banana has compared to the other objects. On the other hand, the car toy was relatively the hardest object to recognize across most of the examined techniques. However, its recognition was significantly enhanced when using all enhancement techniques compared to, for example, the control group (Control - Average Time Taken: 120 sec, Recognition Accuracy: 0%, Average Confidence: 1; All enhancement - Average Time Taken: 70.8 sec, Recognition Accuracy: 80%, Average Confidence: 4.4). In general, there was no object that was hard to recognize when using all enhancement techniques combined, which indicates that the proposed approach enhances the perception of these objects.



(a)



(b)



(c)

Figure 6.6: Single Objects Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. * $P < 0.05$, ** $P < 1e - 04$, two-sample t-test.

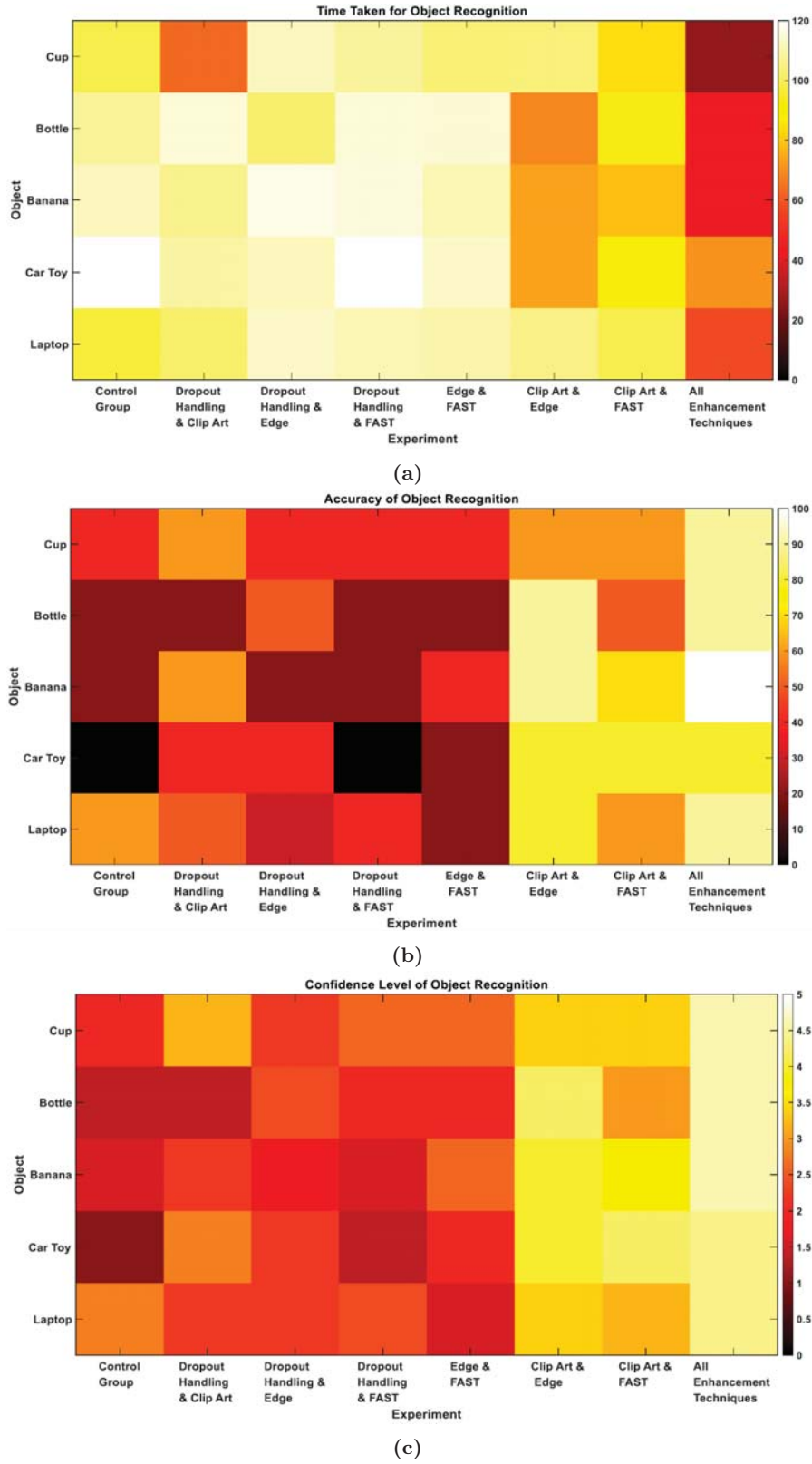
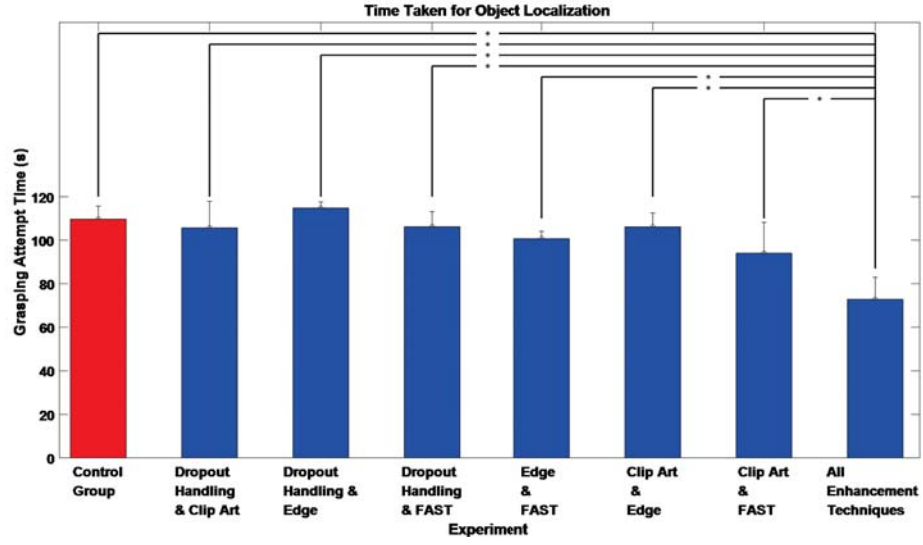


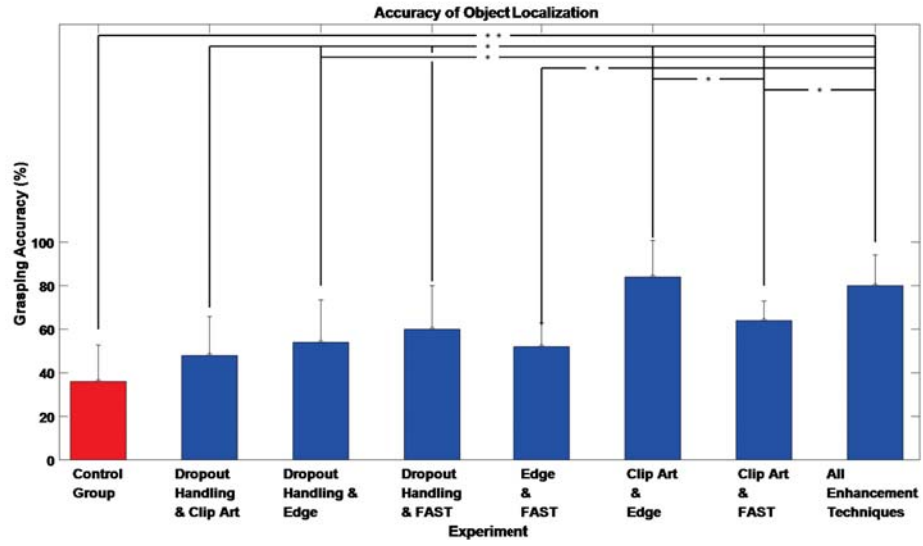
Figure 6.7: Single Objects Recognition Performance for Each Displayed Object. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level.

We also examined the ability of the subjects to localize the objects using the proposed enhancements. Figure 6.8a shows the grasping attempt time of the participants demonstrating that the least time taken for the attempt to grasp an object is the experiment with all the enhancement techniques applied compared to the other experiments (Control Group: 109.68 ± 5.98 sec, Dropout Handling & Clip Art: 105.68 ± 12.15 sec, Dropout Handling & Edge Enhancement: 114.8 ± 2.8 sec, Dropout Handling & FAST: 106.2 ± 6.94 sec, Edge Enhancement & FAST: 100.72 ± 3.38 sec, Clip Art & Edge Enhancement: 106.16 ± 6.29 sec, Clip Art & FAST: 94.04 ± 14.17 sec, and All Enhancement Techniques: 72.84 ± 10.13 sec). In addition, Figure 6.8b shows the accuracy of the participants in correctly grasping the object. The figure demonstrates that the experiments that contain clip art and edge enhancement (experiments 6 and 8), resulted in the highest grasping accuracy (Control Group: $36 \pm 16.73\%$, Dropout Handling & Clip Art: $48 \pm 17.89\%$, Dropout Handling & Edge Enhancement: $52 \pm 17.89\%$, Dropout Handling & FAST: $60 \pm 20\%$, Edge Enhancement & FAST: $52 \pm 10.95\%$, Clip Art & Edge Enhancement: $84 \pm 16.73\%$, Clip Art & FAST: $64 \pm 8.94\%$ and All Enhancement Techniques: $80 \pm 14.14\%$). The results indicate that the proposed enhancement techniques preserved and simplified the details that describe the identity of the object.

We finally assessed the performance achieved using different approaches with respect to the ability of the subjects to grasp each presented object. Figure 6.9 demonstrates that the car toy was the easiest to grasp compared to other objects, achieving an average grasping accuracy of 100% when all enhancement techniques were applied. Based on feedback from the subjects, the car toy was the easiest to grasp because of the wheels as they have a distinct circular shape that was easily located. On the other hand, despite the banana being the easiest to recognize, it was shown to be relatively the hardest to grasp. However, when all enhancement techniques were applied, an elevated grasping accuracy of 80% was achieved. This could be attributed to the fact that the banana was placed flat on the table, making it harder for the subjects to grasp. The figure also indicates that there is no direct correlation between the grasping attempt time and the grasping accuracy, which could be explained given the definition of the grasping attempt time that does not necessarily map to a correct grasp.



(a)



(b)

Figure 6.8: Single Objects Experimental Results Statistics. (a) Grasping Attempt Time. (b) Grasping Accuracy. $*P < 0.05$, $**P < 1e - 04$, two-sample t-test.

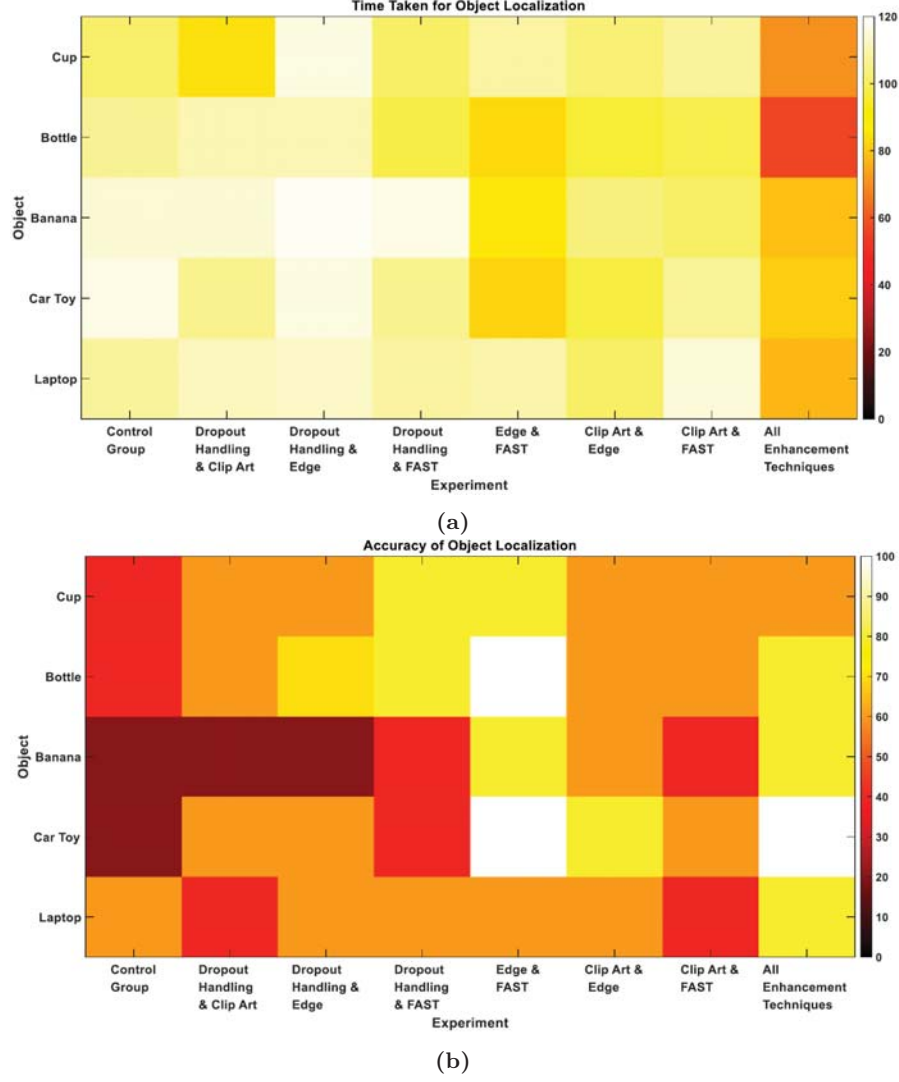


Figure 6.9: Single Objects Grasping Performance for Each Displayed Object. (a) Grasping Attempt Time. (b) Grasping Accuracy.

6.3.2.2 Multiple Objects' Experiments

In the second set of experiments, we examined the proposed approach in a more complex scene that contains multiple objects (a pair of presented objects) as opposed to single objects. Only two setups were examined: “Control Group” and “All Enhancement Techniques”, to compare the results obtained if no enhancement technique was used (i.e., the original real object utilized) to using all enhancement techniques in the phosphene simulation. The setup with “All Enhancement Techniques” was used since it resulted in the best performance in terms of object recognition and localization as observed in the single objects experiments. In terms of object recognition, Figure 6.10a shows the time taken to correctly recognize the objects in the scene illustrating that the experiment with all enhancement techniques gave shorter period of time in objects recognition unlike that of the control group experiment (Control Group: 228.33 ± 14.01 sec and All Enhancement Techniques: 113.93 ± 30.38 sec). Figure 6.10b shows the accuracy of the correctly recognized objects out of the total objects.

The figure shows that the experiment with all enhancement techniques also gave higher accuracy compared to that of the control group experiment (Control Group: $30 \pm 10\%$ and All Enhancement Techniques: $76.67 \pm 11.55\%$). Finally, Figure 6.10c shows the confidence level of the participants, where the experiment with all the enhancement techniques resulted in higher confidence level compared to that of the control group experiment (Control Group: 1.87 ± 0.12 and All Enhancement Techniques: 4 ± 0.53). For object localization, Figure 6.10d shows the time taken for grasping the objects correctly. The figure demonstrates, consistent with the object recognition experiments, that the participants of the experiment in which all the enhancement techniques were applied needed less amount of time for grasping the objects compared to the control group experiment (Control Group: 221.6 ± 22.2 sec and All Enhancement Techniques: 133.93 ± 31.49 sec). Figure 6.10e shows the participants accuracy in correctly grasping the objects, indicating higher accuracy when all the enhancement techniques are applied compared to that of the control group experiment (Control Group: $20 \pm 20\%$ and All Enhancement Techniques: $66.67 \pm 11.55\%$).

We finally evaluated the performance per each pair of presented objects. Figure 6.11 illustrates the outcome of each of the evaluation metrics for each object within each pair. The figure demonstrates that the recognition of the bottle and the cup was the easiest when they were presented together (Recognition Accuracy: 90% for both objects). Consistent results were also obtained for the localization of both objects compared to other objects (Localization Accuracy: 80%). This could be attributed to the significant difference in the shape between the two objects.

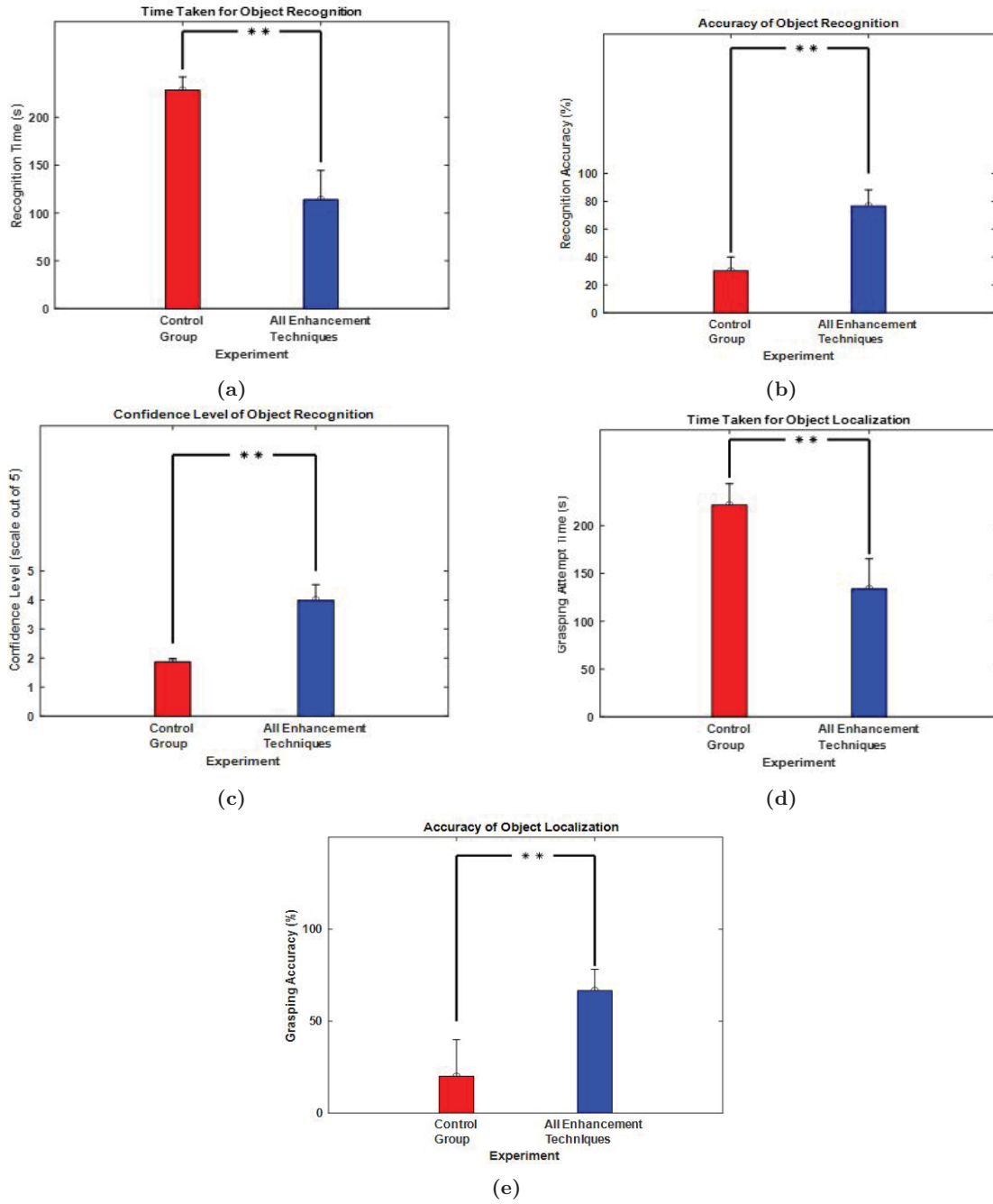


Figure 6.10: Multiple Objects Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. ** $P < 1e - 04$, two-sample t-test.

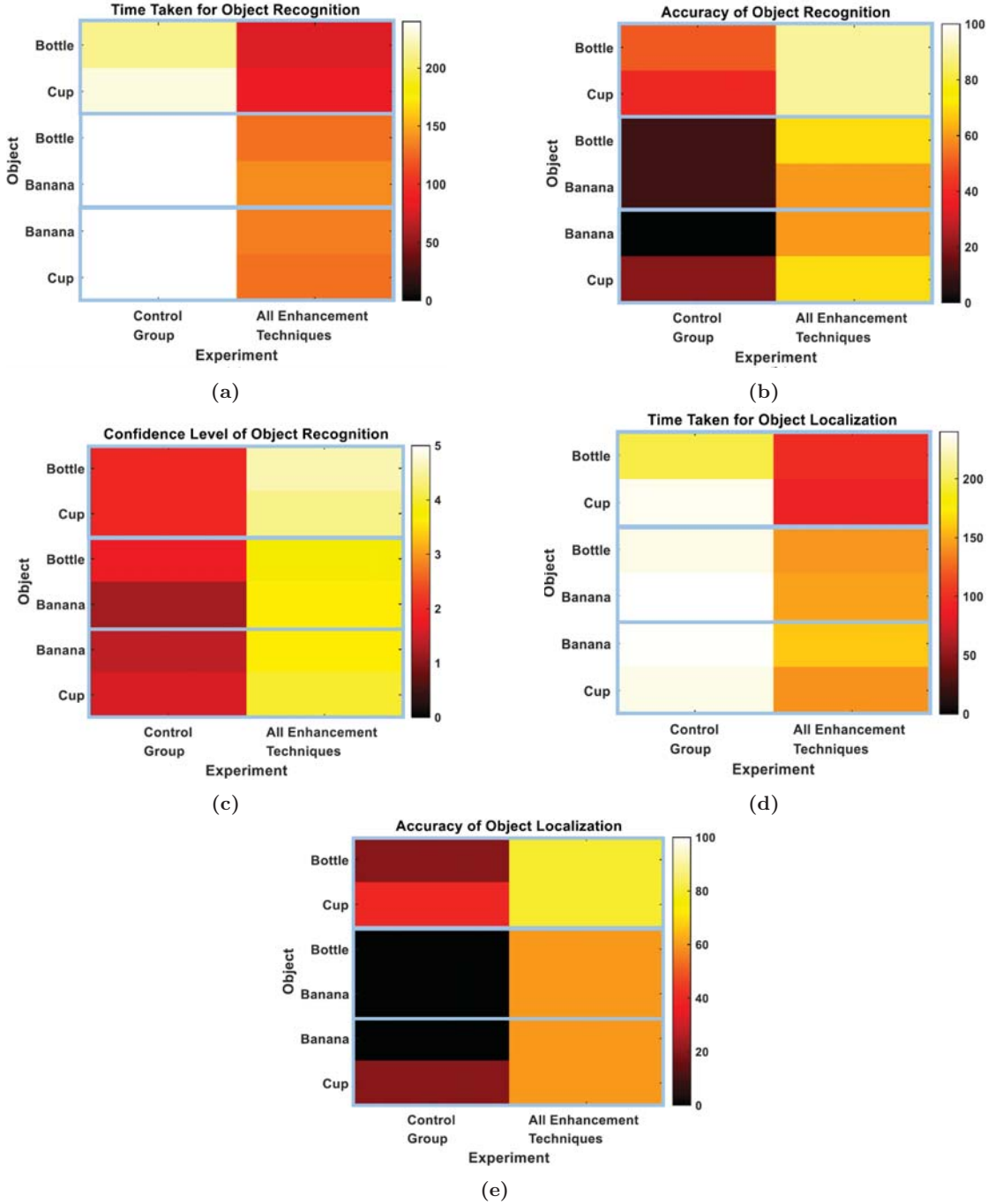


Figure 6.11: Multiple Objects Results for Each Object. (a) Recognition Time. (b) Recognition Accuracy (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. Blue Rectangles in All Figures Represent the Pairs of Objects that were Presented Together.

6.3.2.3 Navigation Experiments

In the last set of experiments, we examined the ability of the participants to navigate freely in order to recognize and grasp an object. In terms of object recognition, Figure 6.12a shows the time taken to correctly recognize objects. Consistent with the previous experiments, the experiment with all the enhancement techniques resulted in less time to recognize the objects compared to that of the control group experiment (Control Group: 126.4 ± 25.67 sec and All

Enhancement Techniques: 63.33 ± 17.81 sec). Figure 6.12b confirmed the same conclusion, where higher recognition accuracy was achieved when all the enhancement techniques were applied compared to that of the control group experiments (Control Group: $26.67 \pm 11.55\%$ and All Enhancement Techniques: $76.67 \pm 15.28\%$). Figure 6.12c shows a similar result when measuring the confidence level of the participants (Control Group: 2.13 ± 0.42 and All Enhancement Techniques: 4.13 ± 0.64).

To demonstrate the results of the participants in object localization in this setup, Figure 6.12d and Figure 6.12e show the time taken and the accuracy achieved by the participants to correctly grasp the object, respectively. Consistent with previous results, less amount of time and higher accuracy are achieved when using all the enhancement techniques compared to that of the control group experiments (Time taken: Control Group: 111.8 ± 5.63 sec and All Enhancement Techniques: 75.27 ± 15.45 sec; Accuracy: Control Group: $60 \pm 20\%$ and All Enhancement Techniques: $93.33 \pm 11.55\%$).

Finally, to assess the performance for each object, Figure 6.13 demonstrates the performance per each presented object. The figure shows that the recognition and localization of the bottle and the chair was relatively easier than those of the backpack. The chair and the bottle were easier to recognize and localize since the former has a very distinct shape with 4 legs for the chair while the latter has a simple shape of a vertical cylinder that could be easily recognized.

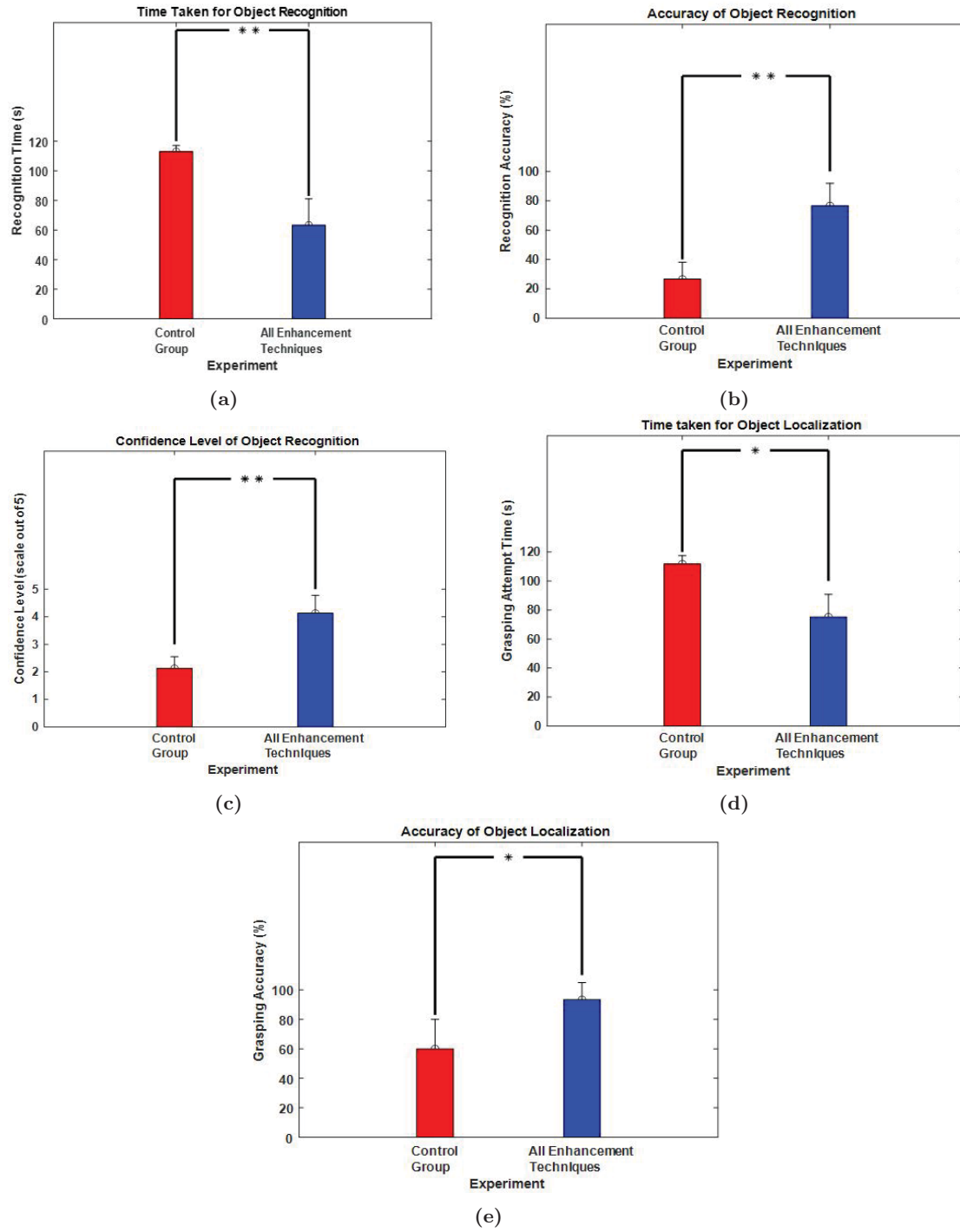


Figure 6.12: Navigation Experimental Results Statistics. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy. * $P < 0.05$, ** $P < 1e - 04$, two-sample t-test.

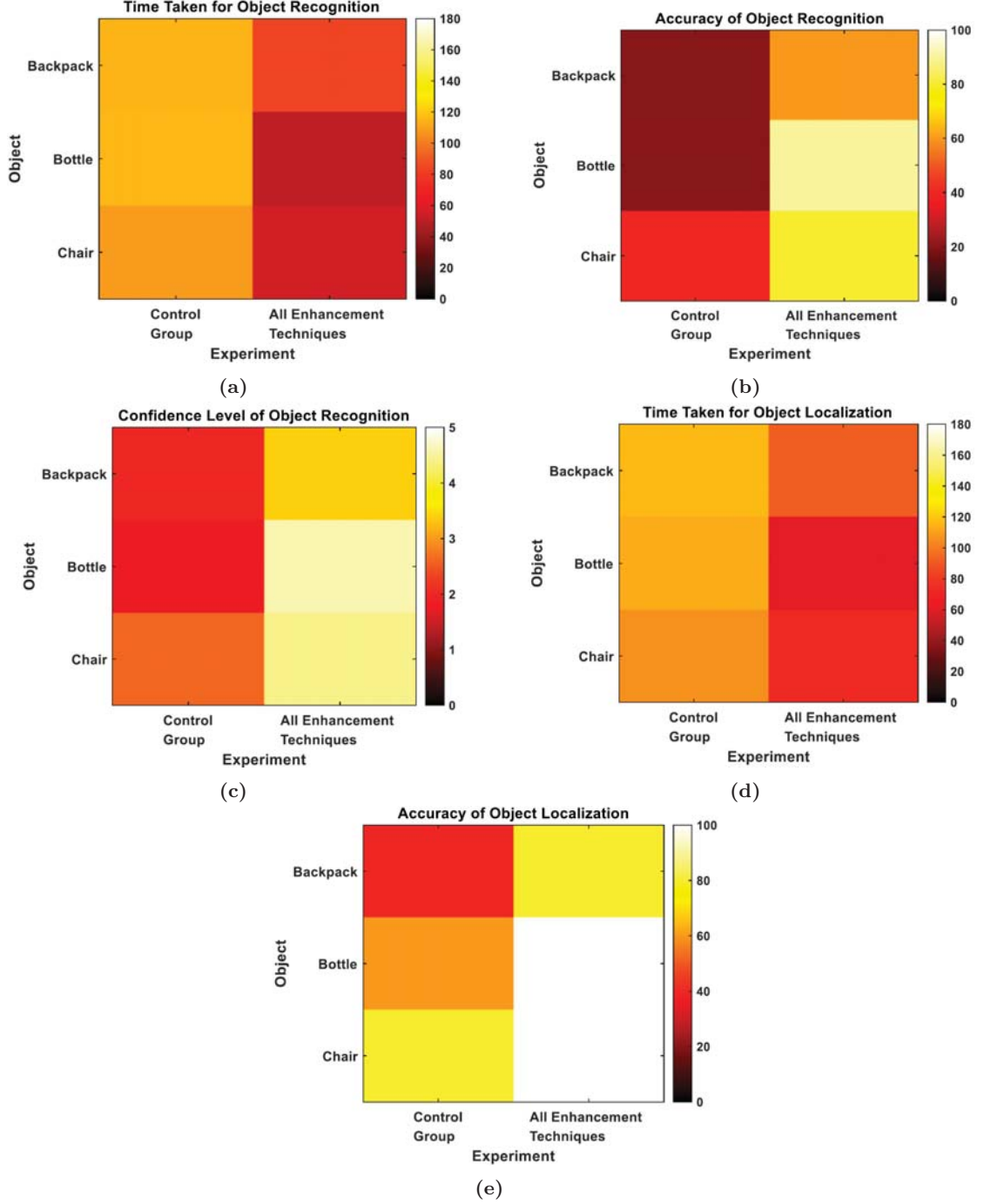


Figure 6.13: Navigation Results for Each Object. (a) Recognition Time. (b) Recognition Accuracy. (c) Confidence Level. (d) Grasping Attempt Time. (e) Grasping Accuracy.

6.4 Analysis of the Utilization of the Enhancement Techniques

We proposed a system for enhancing object recognition and localization through real-time mixed reality simulation. The twelve experiments conducted on the corrected/normally sighted participants indicate that using clip art in place of the real object image significantly enhances the recognition of objects and that using edge enhancement and FAST corner detection significantly enhances objects localization. This could be attributed to the object simplification

provided by using clip art and detail preservation provided by edge sharpening and FAST corner detection. Furthermore, the usage of the dropout handling approach enabled clearer representation of the objects giving the chance for accurate recognition of objects. The multiple objects' experiments show the effect of using YOLO for objects' detection to get the labels of the objects to retrieve the corresponding clip art representation. This enabled accurate selection of clip art object that identifies the real object's identity. It is noteworthy that using clip art as opposed to providing an audio description of the viewed object is in alignment with the purpose of visual prostheses to partially restore vision without relying on other senses. Additionally, using clip art could be superior to using a textual description of the viewed object as this will require more processing time to display the detected word character by character. This is because it might be hard to recognize as a whole word through the typical limited resolution of visual prostheses. In addition, this might be problematic in case of having a long word since a real patient will not be able to memorize the sequence of letters they have already seen.

Previous studies focused on objects segmentation to retrieve each object, either in an image with multiple objects or a single object, at a time and display it in phosphene simulation. However, these segmented objects will remain with their details that will not be obvious when displayed in the low-resolution environment of the phosphene simulation [159]. This was solved in our proposed approach where the usage of clip art that includes abstract representation of any object enabled better visualization and recognition of objects when displayed in the low-resolution environment of the phosphene simulation. Moreover, contrast enhancement was used in previous studies to enhance object recognition, which might not help in providing a clear representation of the objects especially when the scene is a complex one [60]. Furthermore, wavelet-based image processing techniques were addressed to enhance the recognition in the low-resolution environment [160]. However, this might also not provide clear representation of the images since the objects remain with their details. Finally, while edge enhancement techniques were examined before examined [150], they were not combined with using clip art representation or dropout handling.

The navigation experiments show that allowing the participants to navigate while wearing the VR headset in real-time and perceiving the phosphene simulation of the physical floor, eases the ability to indicate that an object is placed on the floor and thus, the object localization is done easily. However, there could be some difficulty in recognizing the object due to the shadows of the light reflections in an indoor scene so the introduction of clip art with edge enhancement and FAST along with translating it to the location with minimum number of dropout locations, will facilitate the ability of the participants to correctly recognize an object. On the other hand, contour-based scene simplification via mobility was used in a real-world indoor scene using simulated prosthetic vision, where the representation of just the outline of an arbitrary object in addition to the discontinuities that are present in the contours hindered recognition accuracy, unlike the usage of clip art along with the other proposed enhancement techniques [160].

6.5 Conclusion

We proposed an image processing approach examined in real-time mixed reality simulation to enhance vision perceived via visual prostheses. We introduced four enhancement techniques which are clip art representation, edge enhancement, FAST and dropout handling. Twelve experiments were conducted in real-time in MR to measure the ability of object recognition and localization. Three metrics were used in object recognition evaluation which were recognition time, recognition accuracy and confidence level. Moreover, two metrics were used to measure the object localization which were grasping attempt time denoting the attempt to localize the object and the grasping accuracy denoting the correctness of localization of the objects without using the sense of touch. The results demonstrate that the introduction of the four enhancement techniques gave the highest recognition accuracy, confidence level and grasping accuracy along with the least recognition time and grasping time.

7

Conclusion and Future Work

7.1 Conclusion

Visual prostheses have recently demonstrated success in partial restoration of vision, giving hope to blind people. This could enable blind people to regain their confidence and independence in performing daily activities. Apart from this partial restoration of vision, some challenges arose which hinder the quality of the perceived image. Some of these challenges are the limited visual field, the limited number of electrodes, the limited number of gray levels and the low spatial resolution of the perceived image in a typical visual prosthetic device. In this thesis, we proposed diverse techniques to solve the difficulty faced by visual prosthetic users in recognizing and localizing the objects correctly. First, a technique was proposed to translate the object of interest to a place in the visual field that includes the minimum amount of electrode dropouts. For the largest percentage value added as an electrode dropout in this thesis, which is 30%, the recognition time of the dropout handling experiment was half that of the no dropout experiment, showing the efficacy of the proposed dropout handling technique. For the recognition accuracy of the dropout handling technique, it recorded 13% more than that of the recognition accuracy of the no dropout handling technique. Second, a YOLO-based approach was used to get the clip art representation of the detected object using Google Images and represent this clip art in phosphene simulation, simplifying the presented scene, leading to easier and faster object recognition. The clip art utilization resulted in a recognition accuracy 70% more than that of utilization of the real images. The recognition time of clip art images was 6 sec less than that of the utilization of the real images, showing the efficacy of the utilization of clip art in visual prostheses. Third, a GAN-based deep learning model was proposed to generate the clip art image of a given input image. Fourth, we proposed three additional enhancement steps, combined with the GAN-generated clip art, to allow better object recognition and localization. A number of experiments was conducted to measure the performance of a normally/correctly sighted participants in the ability to correctly recognize and localize objects. Some of the experiments were performed using a computer screen and others were conducted using mixed reality. The results demonstrate the efficacy

of our proposed techniques showing that our proposed methodologies outperformed those available in literature.

7.2 Future Work

Despite the success of the approaches proposed in this thesis, further extensions can be performed to enhance the perception of images for visual prostheses users. While the results achieved from PVGAN using simulated prosthetic vision indicate the significance of using clip art representation as input to visual prostheses, it remains to be tested on real patients. Feedback from patients would enable better tuning of the model. In addition, using PVGAN in real-time could be examined to measure the amount of delay caused by PVGAN in generating clip arts. Using GANs and deep learning in general is known to induce some delays [161]. It remains to be tested if such delay would have an impact on visual prostheses users' experience.

This work could be also further extended by examining other GAN architectures in addition to testing it on implanted patients. For instance, Spatial Attention GAN model (SPA-GAN) can be tested since it could produce realistic output images, where SPA-GAN computes the attention in its discriminator and uses it to assist the generator in concentrating more on the most discriminative regions between the source and target domains [162]. Moreover, DualGAN, with reference to dual learning from natural language translation, can be utilized, where image translators can be trained on two sets of unlabeled photos from different domains. While the DualGAN learns to reverse the task, the primal GAN acquires the ability to translate images from the domain X to the domain Y . Primal and dual tasks create a closed loop that enables translation and reconstruction of images from either domain. DualGAN looks similar to that of CycleGAN in terms of the main functionality, however, the GANs utilized in DualGAN and CycleGAN are Least square GAN and WGAN (i.e., Wasserstein GAN), respectively [163]. In addition, Contrast-GAN model incorporates an adversarial distance comparison objective, for the purpose of optimising one conditional generator and numerous semantically aware discriminators, which can be used in image-to-image translation [164]. Additionally, Multimodal Unsupervised Image-to-Image Translation (MUNIT) can be utilized, which implies that photos from diverse domains have a common content space, but not a common style space [165]. Finally, a ganimorph model can be utilized to add dilated convolutions to the discriminator architecture. Then, the discriminator output thus makes it easier for the generator to transfer more precise information from the discriminator [166].

Although YOLO was capable of detecting the objects utilized in the experiments, YOLO can be retrained to cover a wider variety of other objects, both non-rigid and rigid objects, to be able to detect any object that a visual prostheses user might encounter. YOLO retraining has been proven to be efficient and has been able to provide remarkable accuracies when retrained using non-rigid objects such as clothing [167], protective clothing worn by workers to ensure their safety [168], hair and the upper body part of people [169]. Moreover, displaying the clip art in 3D and matching its perspective to the viewed object could allow better visualization of the orientation of the actual object.

While the achieved results from the MR simulation experiments, demonstrate the significance of the proposed enhancement techniques, further modifications could be performed.

For instance, the indoor room utilized in the MR experiments included other objects that were not part of the experiment, which sometimes confused some of the participants, thinking that those objects belong to the conducted experiment. In addition, in the multiple objects' experiments, the utilization of more than two objects at a time could be performed to measure the recognition and localization accuracy when having a large number of objects that mimics a real-life perceived scene. The proposed enhancement techniques could be examined in an outdoor environment to measure the efficacy the proposed work and then examined in implanted visual prostheses users.

Our proposed image enhancement techniques can be used in future visual prosthetic systems as it is expected to provide high recognition ability to implanted patients as a result of using the clip art simple representation of objects. Moreover, for future visual prosthetic systems, the number of electrodes is predicted to be larger than the current number of electrodes which could reach 32×32 enabling better perception of the images. One way that our system can be used in real visual prosthesis system is by displaying the real scene followed by the object of interest lightened up and all other objects darkened to enable object localization. This could be followed by the display of the clip art representation of that object. All the objects of interests in the scene will be treated in the same manner until all the clip art representations for all the objects are displayed in order.

A limitation in visual prosthetic devices is the limited number of electrodes in the implant, which results in a low spatial resolution, that hinders the full perception of the visual scene [41]. Such limitation could be alleviated by increasing the number of electrodes, which is expected to enhance the spatial resolution of the images perceived via visual prostheses [42, 170]. However, the perceived image will remain far from the image perceived by normally sighted individuals. Another limitation is the limited number of gray levels available through these devices to represent the perceived image, which affects preserving the details in the image [42]. This could be solved via increasing some system-level constraints and restrictions such as wireless transmission bandwidth and processing capabilities of the implanted module so that the number of gray levels can be increased [21, 171]. In addition, electrodes malfunctioning might happen over time causing dropouts at the corresponding location in the visual field [43]. This dropout can be handled using the proposed approach in this thesis, by translating the object of interest to a location in the visual field that contains the minimum number of dropouts for better visualization and recognition [151]. In the future, other image processing and computer vision techniques, such as correlation, could be utilized to solve the loss of information from an image due to the dropouts.

Despite that our simulation followed the retinal implant, all the proposed approaches in this thesis are not restricted to only retinal implants but they can be utilized in other types including optic nerve, thalamic and cortical ones. Finally, our proposed approaches can be also examined for efficacy in optogenetics visual prosthesis, which is a totally different visual prosthesis compared to the traditional ones.

References

- [1] *World Health Organization Blindness and Visual Impairment Record*. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>. Accessed: 2022-08-08.
- [2] *BlindLook Blindness Record*. <https://www.blindlook.com/blog/detail/the-population-of-blind-people-in-the-world>. Accessed: 2022-08-08.
- [3] Aline Darc Piculo dos Santos et al. “Are electronic white canes better than traditional canes? A comparative study with blind and blindfolded participants”. In: *Universal Access in the Information Society* 20.1 (2021), pp. 93–103.
- [4] Mukesh Prasad Agrawal and Atma Ram Gupta. “Smart stick for the blind and visually impaired people”. In: *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE. 2018, pp. 542–545.
- [5] Gerard Lacey and Kenneth M Dawson-Howe. “The application of robotics to a mobility aid for the elderly blind”. In: *Robotics and Autonomous Systems* 23.4 (1998), pp. 245–252.
- [6] Robert Evan Ornstein and Richard F Thompson. *The amazing brain*. Houghton Mifflin Harcourt, 1986.
- [7] Margaret Wong-Riley. “Energy metabolism of the visual system”. In: *Eye and brain* 2 (2010), p. 99.
- [8] Jiawei Zhang. “Secrets of the brain: an introduction to the brain anatomical structure and biological function”. In: *arXiv preprint arXiv:1906.03314* (2019).
- [9] Göran Darius Hildebrand and Alistair R Fielder. “Anatomy and physiology of the retina”. In: *Pediatric retina*. Springer, 2011, pp. 39–65.
- [10] Ethan D Cohen. “Prosthetic interfaces with the visual system: biological issues”. In: *Journal of neural engineering* 4.2 (2007), R14.
- [11] Thomas Euler et al. “Retinal bipolar cells: elementary building blocks of vision”. In: *Nature Reviews Neuroscience* 15.8 (2014), pp. 507–519.
- [12] Richard H Masland. “The fundamental plan of the retina”. In: *Nature neuroscience* 4.9 (2001), pp. 877–886.
- [13] Diego Ghezzi. “Translation of a photovoltaic retinal prosthesis”. In: *Nature Biomedical Engineering* 4.2 (2020), pp. 137–138.
- [14] Eduardo Fernandez and Klaus-Peter Hoffmann. “Visual prostheses”. In: *Springer handbook of medical technology*. Springer, 2011, pp. 821–834.

REFERENCES

- [15] Eliza Strickland and Mark Harris. “What Happens When a Bionic Body Part Becomes Obsolete?: Blind People with Second Sight’s Retinal Implants Found Out”. In: *IEEE Spectrum* 59.3 (2022), pp. 24–31.
- [16] Gerald J Chader, James Weiland, and Mark S Humayun. “Artificial vision: needs, functioning, and testing of a retinal electronic prosthesis”. In: *Progress in brain research* 175 (2009), pp. 317–332.
- [17] Stuart L Fine et al. “Age-related macular degeneration”. In: *New England Journal of Medicine* 342.7 (2000), pp. 483–492.
- [18] Cordelia Erickson-Davis and Helma Korzybska. “What do blind people “see” with retinal prostheses? Observations and qualitative reports of epiretinal implant users”. In: *PloS one* 16.2 (2021), e0229189.
- [19] Gislin Dagnelie et al. “Performance of real-world functional vision tasks by blind subjects improves after implantation with the Argus® II retinal prosthesis system”. In: *Clinical & experimental ophthalmology* 45.2 (2017), pp. 152–159.
- [20] Robert K Shepherd et al. “Visual prostheses for the blind”. In: *Trends in biotechnology* 31.10 (2013), pp. 562–571.
- [21] Mohammad Hossein Maghami et al. “Visual prostheses: The enabling technology to give sight to the blind”. In: *Journal of ophthalmic & vision research* 9.4 (2014), p. 494.
- [22] Samir Damle et al. “Vertically integrated photo junction-field-effect transistor pixels for retinal prosthesis”. In: *Biomedical Optics Express* 11.1 (2020), pp. 55–67.
- [23] 28 IEEE Standards Coordinating Committee et al. “IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic fields, 3kHz to 300GHz”. In: *IEEE C95. 1-1991* (1992).
- [24] Armin Najarpour Foroushani, Christopher C Pack, and Mohamad Sawan. “Cortical visual prostheses: from microstimulation to functional percept”. In: *Journal of neural engineering* 15.2 (2018), p. 021005.
- [25] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Jose J Guerrero. “Structural and object detection for phosphene images”. In: *arXiv preprint arXiv:1809.09607* (2018).
- [26] NR Srivastava et al. “Estimating phosphene maps for psychophysical experiments used in testing a cortical visual prosthesis device”. In: *2007 3rd International IEEE/EMBS Conference on Neural Engineering*. IEEE. 2007, pp. 130–133.
- [27] Ying Zhao et al. “Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision”. In: *Information Sciences* 180.16 (2010), pp. 2915–2924.
- [28] Spencer C Chen et al. “Simulating prosthetic vision: I. Visual models of phosphenes”. In: *Vision research* 49.12 (2009), pp. 1493–1506.
- [29] SC Chen et al. “Visual acuity measurement of prosthetic vision: a virtual-reality simulation study”. In: *Journal of neural engineering* 2.1 (2005), S135.

-
- [30] Victor Vergnienx, Marc J-M Macé, and Christophe Jouffrais. “Spatial navigation with a simulated prosthetic vision in a virtual environment”. In: *Workshop Neuro-Comp/KEOpS’12*. 2012.
 - [31] Jacob Granley and Michael Beyeler. “A Computational Model of Phosphene Appearance for Epiretinal Prostheses”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 4477–4481.
 - [32] Peng Xia, Jie Hu, and Yinghong Peng. “Adaptation to phosphene parameters based on multi-object recognition using simulated prosthetic vision”. In: *Artificial Organs* 39.12 (2015), pp. 1038–1045.
 - [33] Tai-Chi Lin et al. “Retinal prostheses in degenerative retinal diseases”. In: *Journal of the Chinese Medical Association* 78.9 (2015), pp. 501–505.
 - [34] Ajay Banarji et al. “Visual prosthesis: Artificial vision”. In: *Medical Journal, Armed Forces India* 65.4 (2009), p. 348.
 - [35] Hieu T Nguyen et al. “Thalamic visual prosthesis”. In: *IEEE Transactions on Biomedical Engineering* 63.8 (2016), pp. 1573–1580.
 - [36] Léo Pio-Lopez, Romanos Poulkouras, and Damien Depannemaecker. “Visual cortical prosthesis: an electrical perspective”. In: *Journal of Medical Engineering & Technology* 45.5 (2021), pp. 394–407.
 - [37] Angelica Perez Fornos. “Minimum requirements for a retinal prosthesis to restore useful vision”. PhD thesis. University of Geneva, 2006.
 - [38] John Martin Barrett, Rolando Berlinguer-Palmini, and Patrick Degenaar. “Optogenetic approaches to retinal prosthesis”. In: *Visual neuroscience* 31.4-5 (2014), pp. 345–354.
 - [39] Lauren N Ayton et al. “An update on retinal prostheses”. In: *Clinical Neurophysiology* 131.6 (2020), pp. 1383–1398.
 - [40] JR Boyle, AJ Maeder, and WW Boles. “Image enhancement for electronic visual prostheses”. In: *Australasian Physics & Engineering Sciences in Medicine* 25.2 (2002), pp. 81–86.
 - [41] Eduardo Fernandez. “Development of visual Neuroprostheses: trends and challenges”. In: *Bioelectronic medicine* 4.1 (2018), pp. 1–8.
 - [42] Alejandro Barriga-Rivera et al. “Visual prosthesis: interfacing stimulating electrodes with retinal neurons to restore vision”. In: *Frontiers in neuroscience* 11 (2017), p. 620.
 - [43] Yanyu Lu et al. “Recognition of objects in simulated irregular phosphene maps for an epiretinal prosthesis”. In: *Artificial organs* 38.2 (2014), E10–E20.
 - [44] Jessica L Irons et al. “Face identity recognition in simulated prosthetic vision is poorer than previously reported and can be improved by caricaturing”. In: *Vision research* 137 (2017), pp. 61–79.
 - [45] Carlo Tomasi. “Histograms of oriented gradients”. In: *Computer Vision Sampler* (2012), pp. 1–6.
 - [46] Ta Yang Goh et al. “Performance analysis of image thresholding: Otsu technique”. In: *Measurement* 114 (2018), pp. 298–307.

- [47] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. “Image analysis using mathematical morphology”. In: *IEEE transactions on pattern analysis and machine intelligence* 4 (1987), pp. 532–550.
- [48] Francesco GB De Natale and Giulia Boato. “Detecting morphological filtering of binary images”. In: *IEEE Transactions on Information Forensics and Security* 12.5 (2017), pp. 1207–1217.
- [49] Frank Y Shih and O Robert Mitchell. “Skeletonization and distance transformation by greyscale morphology”. In: *Automated Inspection and High-Speed Vision Architectures*. Vol. 849. SPIE. 1988, pp. 80–86.
- [50] Manuel Forero et al. “Mathematical improvement of the Lee’s 3D skeleton algorithm”. In: *Applications of Digital Image Processing XLII*. Vol. 11137. SPIE. 2019, pp. 602–610.
- [51] Zuwairie Ibrahim and Syed Abdul Rahman Al-Attas. “Wavelet-based printed circuit board inspection algorithm”. In: *Integrated Computer-Aided Engineering* 12.2 (2005), pp. 201–213.
- [52] Lanlan Yang et al. “Numerical determination of RVE for heterogeneous geomaterials based on digital image processing technology”. In: *Processes* 7.6 (2019), p. 346.
- [53] Deepak Geetha Viswanathan. “Features from accelerated segment test (fast)”. In: *Proceedings of the 10th workshop on image analysis for multimedia interactive services, London, UK*. 2009, pp. 6–8.
- [54] Edward Rosten, Reid Porter, and Tom Drummond. “Faster and better: A machine learning approach to corner detection”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.1 (2008), pp. 105–119.
- [55] WNJHW Yussof et al. “Performing contrast limited adaptive histogram equalization technique on combined color models for underwater image enhancement”. In: *International Journal of Interactive Digital Media* 1.1 (2013), pp. 1–6.
- [56] Uğur Erkan, Levent Gökrem, and Serdar Enginoğlu. “Different applied median filter in salt and pepper noise”. In: *Computers & Electrical Engineering* 70 (2018), pp. 789–798.
- [57] Jacob Benesty et al. “Study of the Wiener filter for noise reduction”. In: *Speech enhancement*. Springer, 2005, pp. 9–41.
- [58] MH Chang et al. “Facial identification in very low-resolution images simulating prosthetic vision”. In: *Journal of neural engineering* 9.4 (2012), p. 046012.
- [59] Justin R Boyle, Anthony J Maeder, and Wageeh W Boles. “Region-of-interest processing for electronic visual prostheses”. In: *Journal of Electronic Imaging* 17.1 (2008), p. 013002.
- [60] Jing Wang et al. “Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies”. In: *Information Sciences* 277 (2014), pp. 512–524.
- [61] Fei Guo, Yuan Yang, and Yong Gao. “Optimization of visual information presentation for visual prosthesis”. In: *International journal of biomedical imaging* 2018 (2018).
- [62] Rikiya Yamashita et al. “Convolutional neural networks: an overview and application in radiology”. In: *Insights into imaging* 9.4 (2018), pp. 611–629.

-
- [63] Keiron O’Shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).
 - [64] Andrei Dmitri Gavrilov et al. “Preventing model overfitting and underfitting in convolutional neural networks”. In: *International Journal of Software Science and Computational Intelligence (IJSSCI)* 10.4 (2018), pp. 19–28.
 - [65] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).
 - [66] Pranav Adarsh, Pratibha Rathi, and Manoj Kumar. “YOLO v3-Tiny: Object Detection and Recognition using one stage improved model”. In: *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE. 2020, pp. 687–694.
 - [67] Dae-Hwan Kim. “Evaluation of coco validation 2017 dataset with yolov3”. In: *Evaluation* 6.7 (2019), pp. 10356–10360.
 - [68] Yuan Dai et al. “Efficient foreign object detection between PSDs and metro doors via deep neural networks”. In: *IEEE Access* 8 (2020), pp. 46723–46734.
 - [69] Antonia Creswell et al. “Generative adversarial networks: An overview”. In: *IEEE signal processing magazine* 35.1 (2018), pp. 53–65.
 - [70] S Das. “gan architectures you really should know,” in: *Neptune. Ai*. Retrieved October 13 (6), p. 2021.
 - [71] Byeongsu Sim et al. “Optimal transport driven CycleGAN for unsupervised learning in inverse problems”. In: *SIAM Journal on Imaging Sciences* 13.4 (2020), pp. 2281–2306.
 - [72] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.
 - [73] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).
 - [74] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014).
 - [75] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
 - [76] Muyang Li et al. “Gan compression: Efficient architectures for interactive conditional gans”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5284–5294.
 - [77] Amirhossein Tavanaei and Anthony Maida. “BP-STDP: Approximating backpropagation using spike timing dependent plasticity”. In: *Neurocomputing* 330 (2019), pp. 39–47.
 - [78] Huidong Liu, GU Xianfeng, and Dimitris Samaras. “A two-step computation of the exact gan wasserstein distance”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3159–3168.

- [79] Xiaolong Wang and Abhinav Gupta. “Generative image modeling using style and structure adversarial networks”. In: *European conference on computer vision*. Springer. 2016, pp. 318–335.
- [80] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.
- [81] Chenghui Li, Jun Yao, and Tianci Jiang. “Retinal Vessel Segmentation Network Based on Patch-GAN”. In: *Intelligent Life System Modelling, Image Processing and Analysis*. Springer, 2021, pp. 43–53.
- [82] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and José Jesús Guerrero. “Indoor Scenes Understanding for Visual Prosthesis with Fully Convolutional Networks.” In: *Visigrapp (5: Visapp)*. 2019, pp. 218–225.
- [83] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Jose J Guerrero. “Semantic and structural image segmentation for prosthetic vision”. In: *Plos one* 15.1 (2020), e0227677.
- [84] Ying Zhao, AiPing Yu, and DanTong Xu. “Person Recognition Based on FaceNet under Simulated Prosthetic Vision”. In: *Journal of Physics: Conference Series*. Vol. 1437. 1. IOP Publishing. 2020, p. 012012.
- [85] Jaap de Ruyter van Steveninck et al. “End-to-end optimization of prosthetic vision”. In: *Journal of Vision* 22.2 (2022), pp. 20–20.
- [86] Mohammed J Alwazzan, Mohammed A Ismael, and Asmaa N Ahmed. “A hybrid algorithm to enhance colour retinal fundus images using a Wiener filter and CLAHE”. In: *Journal of Digital Imaging* 34.3 (2021), pp. 750–759.
- [87] Arpan Kumar and Anamika Tiwari. “A comparative study of otsu thresholding and k-means algorithm of image segmentation”. In: *Int. J. Eng. Technol. Res* 9 (2019), pp. 2454–4698.
- [88] Stefan Muthers and Andreas Matzarakis. “Use of beanplots in applied climatology—A comparison with boxplots”. In: *Meteorologische Zeitschrift* 19.6 (2010), pp. 641–644.
- [89] Kenji Suzuki et al. “False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network1”. In: *Academic Radiology* 12.2 (2005), pp. 191–201.
- [90] David D Zhou, Jessy D Dorn, and Robert J Greenberg. “The Argus® II retinal prosthesis system: An overview”. In: *2013 IEEE international conference on multimedia and expo workshops (ICMEW)*. IEEE. 2013, pp. 1–6.
- [91] Yvonne Hsu-Lin Luo and Lyndon Da Cruz. “The Argus® II retinal prosthesis system”. In: *Progress in retinal and eye research* 50 (2016), pp. 89–107.
- [92] Jae-Hyun Jung et al. “Active confocal imaging for visual prostheses”. In: *Vision research* 111 (2015), pp. 182–196.
- [93] Katarina Stingl et al. “Interim results of a multicenter trial with the new electronic subretinal implant alpha AMS in 15 patients blind from inherited retinal degenerations”. In: *Frontiers in neuroscience* 11 (2017), p. 445.

-
- [94] Mark S Humayun et al. “Visual perception in a blind subject with a chronic microelectronic retinal prosthesis”. In: *Vision research* 43.24 (2003), pp. 2573–2581.
 - [95] Xiangyang Xu et al. “Characteristic analysis of Otsu threshold and its applications”. In: *Pattern recognition letters* 32.7 (2011), pp. 956–961.
 - [96] Jie Hu et al. “Recognition of similar objects using simulated prosthetic vision”. In: *Artificial organs* 38.2 (2014), pp. 159–167.
 - [97] Laura Ferlauto et al. “Design and validation of a foldable and photovoltaic wide-field epiretinal prosthesis”. In: *Nature communications* 9.1 (2018), pp. 1–15.
 - [98] YH Luo. “Argus® II Retinal Prosthesis System: Clinical & Functional Outcomes”. PhD thesis. UCL (University College London), 2017.
 - [99] Gislin Dagnelie et al. “Real and virtual mobility performance in simulated prosthetic vision”. In: *Journal of neural engineering* 4.1 (2007), S92.
 - [100] Ying Zhao et al. “Chinese character recognition using simulated phosphene maps”. In: *Investigative ophthalmology & visual science* 52.6 (2011), pp. 3404–3412.
 - [101] Yvonne Hsu-Lin Luo et al. “The use of Argus® II retinal prosthesis to identify common objects in blind subjects with outer retinal dystrophies”. In: *Investigative Ophthalmology & Visual Science* 55.13 (2014), pp. 1834–1834.
 - [102] David Avraham et al. “Retinal prosthetic vision simulation: temporal aspects”. In: *Journal of Neural Engineering* 18.4 (2021), p. 0460d9.
 - [103] Jianxin Lin et al. “Conditional image-to-image translation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5524–5532.
 - [104] Saumitra Mishra et al. “Gan-based generation and automatic selection of explanations for neural networks”. In: *arXiv preprint arXiv:1904.09533* (2019).
 - [105] Pan Zhou, Yunqing Hou, and Jiashi Feng. “Deep adversarial subspace clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1596–1604.
 - [106] Anoop Cherian and Alan Sullivan. “Sem-GAN: semantically-consistent image-to-image translation”. In: *2019 IEEE winter conference on applications of computer vision (wacv)*. IEEE. 2019, pp. 1797–1806.
 - [107] Che-Tsung Lin et al. “GAN-based day-to-night image style transfer for nighttime vehicle detection”. In: *IEEE Transactions on Intelligent Transportation Systems* 22.2 (2020), pp. 951–963.
 - [108] Yezhi Shu, Ran Yi, and Yong-Jin Liu. “Cartoon Your Life”. In: *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE. 2021, pp. 1–2.
 - [109] Han Zhang et al. “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5907–5915.
 - [110] Ming-Yu Liu and Oncel Tuzel. “Coupled generative adversarial networks”. In: *Advances in neural information processing systems* 29 (2016).

- [111] Lei Xu et al. “Modeling tabular data using conditional gan”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [112] Emily L Denton, Soumith Chintala, Rob Fergus, et al. “Deep generative image models using a laplacian pyramid of adversarial networks”. In: *Advances in neural information processing systems* 28 (2015).
- [113] Jon Gauthier. “Conditional generative adversarial nets for convolutional face generation”. In: *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester 2014.5* (2014), p. 2.
- [114] Scott Reed et al. “Generative adversarial text to image synthesis”. In: *International conference on machine learning*. PMLR. 2016, pp. 1060–1069.
- [115] Xiaolong Wang and Abhinav Gupta. “Generative image modeling using style and structure adversarial networks”. In: *European conference on computer vision*. Springer. 2016, pp. 318–335.
- [116] Donggeun Yoo et al. “Pixel-level domain transfer”. In: *European conference on computer vision*. Springer. 2016, pp. 517–532.
- [117] Michael Mathieu, Camille Couprie, and Yann LeCun. “Deep multi-scale video prediction beyond mean square error”. In: *arXiv preprint arXiv:1511.05440* (2015).
- [118] Levent Karacan et al. “Learning to generate images of outdoor scenes from attributes and semantic layouts”. In: *arXiv preprint arXiv:1612.00215* (2016).
- [119] Scott Reed et al. “Generating interpretable images with controllable structure”. In: (2016).
- [120] Scott E Reed et al. “Learning what and where to draw”. In: *Advances in neural information processing systems* 29 (2016).
- [121] Yipin Zhou and Tamara L Berg. “Learning temporal transformations from time-lapse videos”. In: *European conference on computer vision*. Springer. 2016, pp. 262–277.
- [122] Deepak Pathak et al. “Context encoders: Feature learning by inpainting”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.
- [123] Jun-Yan Zhu et al. “Generative visual manipulation on the natural image manifold”. In: *European conference on computer vision*. Springer. 2016, pp. 597–613.
- [124] Christian Ledig et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.
- [125] Chuan Li and Michael Wand. “Precomputed real-time texture synthesis with markovian generative adversarial networks”. In: *European conference on computer vision*. Springer. 2016, pp. 702–716.
- [126] Dan Popescu et al. “Retinal blood vessel segmentation using pix2pix gan”. In: *2021 29th Mediterranean Conference on Control and Automation (MED)*. IEEE. 2021, pp. 1173–1178.
- [127] Horea Mureşan and Mihai Oltean. “Fruit recognition from images using deep learning”. In: *arXiv preprint arXiv:1712.00580* (2017).

- [128] Clara Grilo et al. *BRAZIL ROAD-KILL: a data set of wildlife terrestrial vertebrate road-kills*. 2018.
- [129] Yao Yu chen. “Dog and Cat Classification with Deep Residual Network”. In: *Proceedings of the 2020 European Symposium on Software Engineering*. 2020, pp. 137–141.
- [130] Sumanta Bhattacharyya, Arindrajit Seal, and Arindam Mukherjee. “Real-Time Traffic Incidence dataset”. In: *2019 SoutheastCon*. IEEE. 2019, pp. 1–5.
- [131] Tibor Trnovszky et al. “Animal recognition system based on convolutional neural network”. In: *Advances in Electrical and Electronic Engineering* 15.3 (2017), pp. 517–525.
- [132] Horea Mureşan and Mihai Oltean. “Fruit recognition from images using deep learning”. In: *arXiv preprint arXiv:1712.00580* (2017).
- [133] Florian Mormann et al. “A category-specific response to animals in the right human amygdala”. In: *Nature neuroscience* 14.10 (2011), pp. 1247–1249.
- [134] Akash Kumar and Sourya Dipta Das. “Bird species classification using transfer learning with multistage training”. In: *Workshop on Computer Vision Applications*. Springer. 2018, pp. 28–38.
- [135] Omkar M Parkhi et al. “Cats and dogs”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3498–3505.
- [136] K Jemimah. “Recognition of handwritten characters based on deep learning with tensorflow”. In: *Int. Res. J. Eng. Technol.(IRJET)* 6.09 (2019).
- [137] Ahmet Sakir Dokuz. “Fast and efficient discovery of key bike stations in bike sharing systems big datasets”. In: *Expert Systems with Applications* 172 (2021), p. 114659.
- [138] Andrea S Griffin, Daniel T Blumstein, and Christopher S Evans. “Training captive-bred or translocated animals to avoid predators”. In: *Conservation biology* 14.5 (2000), pp. 1317–1326.
- [139] Uğur Erkan, Levent Gökrem, and Serdar Enginoğlu. “Different applied median filter in salt and pepper noise”. In: *Computers & Electrical Engineering* 70 (2018), pp. 789–798.
- [140] Laurent Cabaret, Lionel Lacassagne, and Louiza Oudni. “A review of world’s fastest connected component labeling algorithms: Speed and energy estimation”. In: *Proceedings of the 2014 Conference on Design and Architectures for Signal and Image Processing*. IEEE. 2014, pp. 1–6.
- [141] Sheng He and Lambert Schomaker. “DeepOtsu: Document enhancement and binarization using iterative deep learning”. In: *Pattern recognition* 91 (2019), pp. 379–390.
- [142] Debashis Das. “A minutia detection approach from direct gray-scale fingerprint image using hit-or-miss transformation”. In: *Computational intelligence in pattern recognition*. Springer, 2020, pp. 195–206.
- [143] Yan Song et al. “Medical Image Edge Detection Based on Improved Differential Evolution Algorithm and Prewitt Operator.” In: *Acta Microscopica* 28.1 (2019).
- [144] Rajiv Kapoor et al. “Detection of power quality event using histogram of oriented gradients and support vector machine”. In: *Measurement* 120 (2018), pp. 52–75.

REFERENCES

- [145] Yun Wei et al. “Multi-vehicle detection algorithm through combining Harr and HOG features”. In: *Mathematics and Computers in Simulation* 155 (2019), pp. 130–145.
- [146] Scott H Greenwald et al. “Brightness as a function of current amplitude in human retinal electrical stimulation”. In: *Investigative ophthalmology & visual science* 50.11 (2009), pp. 5017–5025.
- [147] Jacob Thomas Thorn, Enrico Migliorini, and Diego Ghezzi. “Virtual reality simulation of epiretinal stimulation highlights the relevance of the visual angle in prosthetic vision”. In: *Journal of Neural Engineering* 17.5 (2020), p. 056019.
- [148] Michael Beyeler et al. “pulse2percept: A Python-based simulation framework for bionic vision”. In: *BioRxiv* (2017), p. 148015.
- [149] David Avraham and Yitzhak Yitzhaky. “Effects of depth-based object isolation in simulated retinal prosthetic vision”. In: *Symmetry* 13.10 (2021), p. 1763.
- [150] Jason A Dowling, Anthony Maeder, and Wageeh Boles. “Mobility enhancement and assessment for a visual prosthesis”. In: *Medical Imaging 2004: Physiology, Function, and Structure from Medical Images*. Vol. 5369. SPIE. 2004, pp. 780–791.
- [151] Reham H Elnabawy et al. “Electrode Dropout Compensation in Visual Prostheses: An Optimal Object Placement Approach”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 6515–6518.
- [152] Stanislaw Rizzo et al. “The Argus II Retinal Prosthesis: 12-month outcomes from a single-study center”. In: *American journal of ophthalmology* 157.6 (2014), pp. 1282–1290.
- [153] D Nanduri et al. “Retinal prosthesis phosphene shape analysis”. In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2008, pp. 1785–1788.
- [154] Reham H Elnabawy et al. “A YOLO-based Object Simplification Approach for Visual Prostheses”. In: *35th International Symposium on Computer Based Medical Systems (CBMS)*. IEEE. 2022.
- [155] Mikel Val Calvo et al. “Horizon Cyber-Vision: A Cybernetic Approach for a Cortical Visual Prosthesis”. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2022, pp. 380–394.
- [156] Lijun Ding and Ardeshtir Goshtasby. “On the Canny edge detector”. In: *Pattern recognition* 34.3 (2001), pp. 721–725.
- [157] Arpita Mohapatra et al. “Comparative study of corner and feature extractors for real-time object recognition in image processing”. In: *Journal of information and communication convergence engineering* 12.4 (2014), pp. 263–270.
- [158] Edward Rosten and Tom Drummond. “Machine learning for high-speed corner detection”. In: *European conference on computer vision*. Springer. 2006, pp. 430–443.
- [159] Jing Wang et al. “The application of computer vision to visual prosthesis”. In: *Artificial Organs* 45.10 (2021), pp. 1141–1154.

-
- [160] Sheng Li et al. “Image recognition with a limited number of pixels for visual prostheses design”. In: *Artificial organs* 36.3 (2012), pp. 266–274.
 - [161] Muhammad Sarmad, Hyunjoon Jenny Lee, and Young Min Kim. “Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5898–5907.
 - [162] Hajar Emami et al. “Spa-gan: Spatial attention gan for image-to-image translation”. In: *IEEE Transactions on Multimedia* 23 (2020), pp. 391–401.
 - [163] Zili Yi et al. “Dualgan: Unsupervised dual learning for image-to-image translation”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2849–2857.
 - [164] Xiaodan Liang, Hao Zhang, and Eric P Xing. “Generative semantic manipulation with contrasting gan”. In: *arXiv preprint arXiv:1708.00315* (2017).
 - [165] Xun Huang et al. “Multimodal unsupervised image-to-image translation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189.
 - [166] Aaron Gokaslan et al. “Improving shape deformation in unsupervised image-to-image translation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 649–665.
 - [167] Haijun Zhang et al. “ClothingOut: a category-supervised GAN model for clothing segmentation and retrieval”. In: *Neural computing and applications* 32.9 (2020), pp. 4519–4530.
 - [168] Xuanyu Wang et al. “A Safety Helmet and Protective Clothing Detection Method based on Improved-Yolo V3”. In: *2020 Chinese Automation Congress (CAC)*. IEEE. 2020, pp. 5437–5441.
 - [169] Shuo Zhang et al. “Tiny YOLO optimization oriented bus passenger object detection”. In: *Chinese Journal of Electronics* 29.1 (2020), pp. 132–138.
 - [170] Katarina Stingl et al. “Subretinal visual implant alpha IMS–clinical trial interim report”. In: *Vision research* 111 (2015), pp. 149–160.
 - [171] Robert W Thompson et al. “Facial recognition using simulated prosthetic pixelized vision”. In: *Investigative ophthalmology & visual science* 44.11 (2003), pp. 5035–5042.