

Melani Sánchez García

Egocentric Computer Vision and Machine Learning for Simulated Prosthetic Vision

Director/es

Guerrero Campo, José Jesús
Martínez Cantín, Rubén

<http://zaguan.unizar.es/collection/Tesis>

© Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606



Universidad
Zaragoza

Tesis Doctoral

**EGOCENTRIC COMPUTER VISION AND MACHINE
LEARNING FOR SIMULATED PROSTHETIC VISION**

Autor

Melani Sánchez García

Director/es

Guerrero Campo, José Jesús
Martínez Cantín, Rubén

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Ingeniería de Sistemas e Informática

2022



Universidad
Zaragoza

Tesis Doctoral

Egocentric Computer Vision and Machine Learning for Simulated Prosthetic Vision

Autor

Melani Sánchez García

Director/es

GUERRERO CAMPO, JOSE JESUS

MARTINEZ CANTIN, RUBEN

UNIVERSIDAD DE ZARAGOZA

Escuela de Doctorado

Programa de Doctorado en Informática e Ingeniería de Sistemas

2021

Egocentric Computer Vision and Machine Learning for Simulated Prosthetic Vision

Autor

Melani Sánchez García

Director/es

José Jesús Guerrero Campo

Rubén Martínez Cantín

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2021

Acknowledgements

First of all, I MUST thank my supervisors, Josechu Guerrero and Rubén Martínez Cantín, for their help, as well as for trusting me during all these years. Thanks Josechu for encouraging and giving me a chance, valuable advices and think of me as a father saying that: ‘the most important thing is to be happy’. To you Rubén, I will never be able to thank you enough all that you have worked with me during my thesis. Thank you for trying that my work was always impeccable. Thank you for working hard with me every day, even holidays or until late morning hours. It has been a pleasure learning from you. Thanks a million to both of you.

This thesis has been possible thanks to the funding by the Spanish Ministry of Economy, Industry and Competitiveness, with the scholarship that I was conceded (BES-2016-078426), and the projects DPI2015-65962-R and RTI2018-096903-B-I00 (MINECO/FEDER, UE). I’m especially grateful for the opportunities they gave me to attend to conferences and perform abroad stays in two research centers: Centre de Recherche Cerveau & Cognition (CerCo) and Instituto de Biogingeniería, Universidad Miguel Hernández. Apart from the academic value, these stays were amazing personal experiences, thanks to the wonderful people who crossed my path, like Benoit Cottureau and Tushar Chauchan. Thank you Tushar for your dedication. I was very happy working with you.

And I cannot forget the people from Zaragoza. I’ve been here a long time, which has given me the possibility of meeting so many extraordinary people. Some even dare to collaborate with me, like Jesús and Alejandro. And especially you, Emmanuel, for showing up in my way when I needed it most. Honestly, I still don’t know how you did it.

To conclude, I have to dedicate this all to my parents and my sister for their unconditional support over these years. Yes family, sometimes it was hard but it was worth it. Thank you Vicente for being by my side all these years, for making me laugh from the distance and playing with me to PG and of course, for waiting for me in Valencia every weekend.

To my **B**...

Resumen

Las prótesis visuales actuales son capaces de proporcionar percepción visual a personas con cierta ceguera. Sin pasar por la parte dañada del camino visual, la estimulación eléctrica en la retina o en el sistema nervioso provoca percepciones puntuales conocidas como “fosfenos”. Debido a limitaciones fisiológicas y tecnológicas, la información que reciben los pacientes tiene una resolución muy baja y un campo de visión y rango dinámico reducido afectando seriamente la capacidad de la persona para reconocer y navegar en entornos desconocidos. En este contexto, la inclusión de nuevas técnicas de visión por computador es un tema clave activo y abierto. En esta tesis nos centramos especialmente en el problema de desarrollar técnicas para potenciar la información visual que recibe el paciente implantado y proponemos diferentes sistemas de visión protésica simulada para la experimentación. Las contribuciones generales de esta tesis son cinco:

1. Al combinar la salida de dos redes neuronales convolucionales para detectar bordes informativos estructurales y siluetas de objetos, demostramos cómo se pueden reconocer rápidamente diferentes escenas y objetos incluso en las condiciones restringidas de la visión protésica. Nuestro método es muy adecuado para la comprensión de escenas de interiores comparado con los métodos tradicionales de procesamiento de imágenes utilizados en prótesis visuales.
2. Presentamos un nuevo sistema de realidad virtual para entornos de visión protésica simulada más realistas usando escenas panorámicas, lo que nos permitió estudiar sistemáticamente el rendimiento de la búsqueda y reconocimiento de objetos. Las escenas panorámicas permiten que los sujetos se sientan inmersos en la escena al percibir la escena completa (360 grados).
3. Demostramos cómo un sistema de navegación de realidad aumentada para visión protésica ayuda al rendimiento de la navegación al reducir el tiempo y la distancia para alcanzar los objetivos, incluso reduciendo significativamente el número de colisiones de obstáculos. Mediante el uso de un algoritmo de planificación de ruta, el sistema encamina al sujeto a través de una ruta más corta y sin obstáculos.
4. Evaluamos la agudeza visual midiendo la influencia del campo de visión con respecto a la resolución espacial en un nuevo entorno de visión protésica simulada. Nuestro

entorno de simulación aprovecha un software de realidad virtual combinado con una pantalla portátil montada en la cabeza para simular la experiencia de la vida real de usar una prótesis de retina.

5. Proponemos un modelo de Spiking Neural Network que se basa en mecanismos biológicamente plausibles y utiliza un esquema de aprendizaje no supervisado para obtener mejores algoritmos computacionales y mejorar el rendimiento de las prótesis visuales actuales. El modelo propuesto puede hacer uso de la señal de muestreo descendente de la unidad de procesamiento de información de las prótesis retinianas sin pasar por el análisis de imágenes retinianas, proporcionando información útil a los ciegos.

Abstract

Current visual prostheses are capable of providing visual perception to people with certain blindness. Bypassing the damaged part of the visual path, electrical stimulation in the retina or nervous system provokes spot percepts known as “phosphenes”. Due to physiological and technological limitations, the information received by patients has very low resolution, low field of view and reduced dynamic range seriously affecting the person’s ability to recognize and navigate in unknown environments. In this context, the inclusion of new computer vision techniques is an active and open key topic. In this thesis, we particularly focus on the problem of developing techniques to enhance the visual information received by the implanted patient and propose different simulated prosthetic vision systems for experimentation. The overarching contributions of this thesis are fivefold:

1. By combining the output of two Fully Convolutional Networks for structural informative edges and object masks and silhouettes detection, we demonstrated how different scenes and objects can be quickly recognized even under the restricted conditions of prosthetic vision. Our method is well suited for indoor scene understanding over traditional image processing methods used in visual prostheses.
2. We present a new virtual-reality system for more realistic simulated prosthetic vision environments using panoramic scenes, which allowed us to systematically study object search and recognition performance. This system acts as an electronic visual aid that attach to the user’s head and presents information directly to the user’s eyes. The panoramic scenes allow subjects felt immersed in the scene by perceiving the entire scene (360 degrees).
3. We demonstrate how an augmented reality navigation system for prosthetic vision help navigation performance by reducing the time and distance to reach goals, even significantly reducing the number of obstacles collisions. By using a route planning algorithm, the system route the subject through a shorter, obstacle-free route.
4. We assess visual acuity by measuring the influence of the field of view with respect to spatial resolution in a new simulated prosthetic vision environment. Our simulation

environment took advantage of virtual-reality software paired with a portable head-mounted display to simulate the real-life experience of wearing a retinal prosthesis.

5. We propose a Spiking Neural Network model that is based on biologically plausible mechanisms and uses an unsupervised learning scheme to obtain better computational algorithms and improve the performance of current visual prostheses. The proposed model can make use of the down-sampled signal from information processing unit of retinal prostheses bypassing retinal image analysis, providing useful information to the blind.

Contents

1	Introduction	1
1.1	The human visual system	2
1.2	Vision impairments	3
1.3	Sight restoration technologies	4
1.3.1	Principles of prosthetic vision	5
1.3.2	Retinal prostheses	8
1.4	Simulated Prosthetic Vision	10
1.4.1	Visual models of phosphenes	11
1.4.2	Experimentation with SPV	13
1.5	Motivation and challenges	17
1.6	Goals and Contributions	17
1.6.1	Schematic representation of indoor scenes	19
1.6.2	Influence of field of view	19
1.6.3	Augmented reality navigation	20
1.6.4	Visual acuity assessment with VR	21
1.6.5	Spiking neural network	21
1.7	Organization of the Document	22
2	Schematic representation of indoor scenes	24
2.1	Introduction	25
2.2	Methods	27
2.2.1	Subjects	27
2.2.2	Stimuli	28
2.2.3	Experimental setup	34
2.3	Results	38
2.3.1	Comparison of stimuli generation methods	38
2.3.2	Performance analysis of SIE-OMS	40
2.4	Discussion	43
2.5	Conclusions	46
2.6	Related Publications	46
3	Influence of field of view	48
3.1	Introduction	49
3.2	Materials and Methods	52
3.2.1	Participants	53
3.2.2	Simulated Prosthetic Vision (SPV)	53
3.2.3	Procedure	56
3.2.4	Statistical analysis	57
3.3	Results	57
3.4	Discussion	60
3.5	Conclusions	64
3.6	Related Publications	65

4	Augmented reality navigation	66
4.1	Introduction	67
4.2	Methods	70
4.2.1	Subjects	70
4.2.2	Augmented Reality System	71
4.2.3	Procedure	76
4.2.4	Statistical analysis	78
4.3	Results	78
4.4	Discussion	80
4.5	Conclusions	83
4.6	Related Publications	84
5	Visual acuity assessment with VR	85
5.1	Introduction	86
5.2	Methods	88
5.2.1	Participants	89
5.2.2	Simulated Prosthetic Vision (SPV)	89
5.2.3	Procedure	91
5.2.4	Statistical analysis	92
5.3	Results	93
5.3.1	Light perception	93
5.3.2	Time resolution	93
5.3.3	Light location	95
5.3.4	Motion perception	96
5.3.5	Landolt-C orientation	97
5.4	Discussion	97
5.5	Conclusions	101
5.6	Related Publications	101
6	Spiking neural network	102
6.1	Introduction	103
6.2	Spiking Neural Network	105
6.2.1	Spiking Neurons	105
6.2.2	Neural coding	105
6.2.3	Synapses	106
6.2.4	Inhibition	107
6.3	Method	107
6.3.1	Dataset	107
6.3.2	Pre-processing	108
6.3.3	STDP neural network	108
6.3.4	Divisive normalization	112
6.4	Results	112
6.4.1	Classification using supervised or unsupervised network	113
6.4.2	Feature detection	115
6.5	Conclusions	117
6.6	Related Publications	117
7	Conclusions and future prospects	119

8 Summary of results	125
8.1 Research Stays	125
8.2 Supervision of Students	125
8.3 Dissemination	126
8.3.1 Peer-Reviewed Publications	126
8.3.2 Open-Source Software	127
8.3.3 Conference and Research Seminar Attendance	127
References	128
Appendix	152

List of Figures

1.1	Schematic representation of the eye and retina.	2
1.2	Visual impairments	4
1.3	Prosthetic vision	6
1.4	Retinal prostheses.	9
1.5	Simulated Prosthetic Vision (SPV) setups.	10
1.6	Visual moldels of phosphenes.	11
1.7	SPV approaches for different tasks.	14
2.1	Stimuli generation.	29
2.2	Processing pipeline.	30
2.3	Scene layout from an indoor image.	31
2.4	Box and mask branch from OMS.	32
2.5	Objects Masks and Silhouettes (OMS).	33
2.6	SPV and trial setup.	35
2.7	Examples of stimuli used in the experiment.	37
2.8	Global results by phosphenic stimuli method.	39
2.9	Object recognition results for each room-type.	40
2.10	Room identification results for each room-type.	41
2.11	Successful and failed images results.	42
3.1	SPV system.	52
3.2	Data process.	54
3.3	Stimuli conditions in the experiment.	55
3.4	Object classes considered during the experiment.	56
3.5	Trial setup.	58
3.6	Ratio of recognized objects and recognition time.	58
3.7	Ratio of recognized objects and recognition time.	59
3.8	Angular resolution.	61
4.1	Virtual environments used in the experiments.	70
4.2	Guidance methods with door goal.	71
4.3	Guidance methods with bin goal.	72
4.4	Diagram of SPV environment.	73
4.5	SPV system.	76
4.6	Trial setup.	77
4.7	Mean time and distance.	78
4.8	Mean time and distance.	80
4.9	Examples of path trajectories for the three guidance methods.	83
5.1	Data process	89
5.2	Stimuli conditions in the experiment.	91
5.3	Subject and trial setup.	92
5.4	Light perception	94
5.5	Time resolution	94
5.6	Light location	95
5.7	Motion perception	96

5.8	Landolt-C orientation	98
5.9	Visual acuity for Landolt-C.	98
6.1	Spike-timing-dependent-plasticity.	106
6.2	Combination of spatial frequency sub-networks.	109
6.3	Receptive fields learned by different neurons.	113
6.4	Confusion matrix for MNIST and fashion–MNIST.	114
6.5	Feature detection using various spatial frequencies.	115
6.6	Results of features learned by our unsupervised network.	116
A0.1	Vertical lines in indoor environments.	152

List of Tables

2.1	Global object recognition and room identification.	38
2.2	Confusion matrix for room identification based on answered images. . . .	41
2.3	Confusion matrix for room identification based on the total images. . . .	42
3.1	Mean value of the ratio of recognized object and recognition time.	62
4.1	Mean value of time, distance and total number of bumps.	79
5.1	Pixel density on the phosphenic image.	99

Chapter 1

1 Introduction

“The best and most beautiful things in the world cannot be seen or even touched - they must be felt with heart.”
— *Hellen Keller*

How far away is a cure for blindness? Understanding the brain, its functions, and its diseases has motivated researchers in the last century to cross new biomedical frontiers and develop advanced neurotechnologies to help the hundreds of million people worldwide that suffer from neurological disorders [1]. Neuroengineering might nowadays offer potential solutions for neurological afflictions based on prosthetic devices, which establish bidirectional communication between electronic machines and neurons [2–4]. For example, the activity of neurons can be influenced through their excitation or inhibition [5]. Consequently, the modulation of the central and peripheral nervous system allows to induce sensory perception, e.g., for blindness. One approach to stimulate and record neuronal activity is using implantable neural interfaces, which offer an intimate connection with the nervous tissue.

Owing to the incredible recent technological progress in the development of tools and devices aiming at interfacing to the nervous system, considerable research has been done in recent decades to deepen our understanding of the human nervous system or to interact with the nervous system for prosthetic purposes [6–9]. Implantable systems designed and developed to either stimulate the nervous system or directly record neuronal activities are referred to as visual prostheses. In this regard, the proposed thesis is about novel computer vision techniques suitable for improving the visual information received by implanted patients. The work focuses on enhancing important aspects of modern retinal prostheses to obtain a safer and more useful form of artificial vision for people blinded by retinal degeneration diseases. In the following sections, relevant aspects related to the thesis will be introduced; namely, the human visual system, vision impairments,

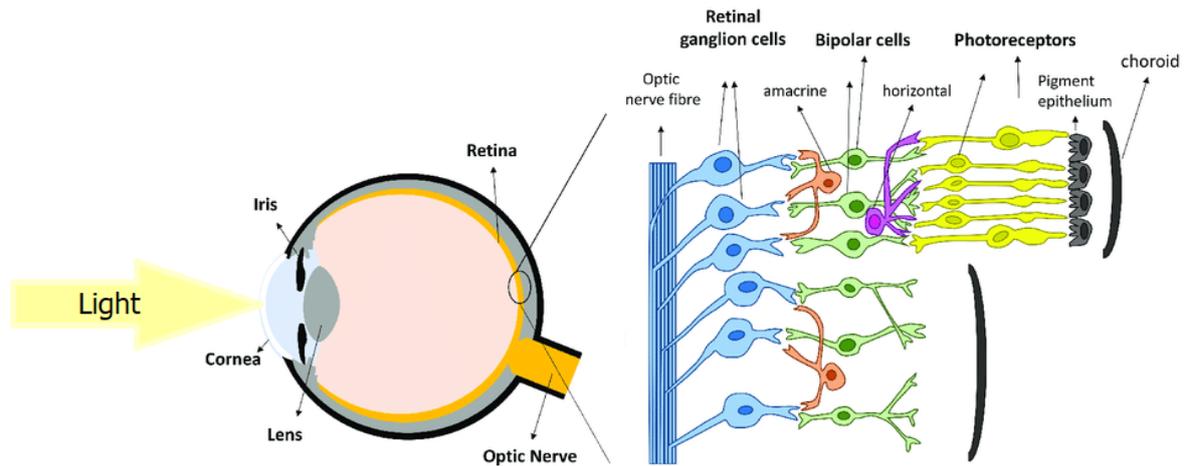


Figure 1.1: Schematic representation of the eye and retina. Light enters the eye through the cornea and is focused by the lens onto the retina. The retina is mainly composed of three layers of neurons, photoreceptors, bipolar cells and retinal ganglion cells, with horizontal and amacrine cells in between. (Adapted from Tong et al. [10]).

some sight restoration technologies (with emphasis on retinal prostheses), and the use of simulated prosthetic vision to improve the design of current visual prostheses, which is the working principle of the research in this thesis.

1.1 The human visual system

Visual perception begins when light enters the eye and induces electrical signals that are subsequently processed within the brain producing an image of the visual world (Figure 1.1). This process starts when the light enters the eye through the cornea, passes the pupil surrounded by the iris, and is focused by the lens onto the retina, the nervous tissue that conveys extraordinary image processing. Light has to travel through the entire thickness of the retina until reaching the outer nuclear layer where photoreceptors (rods and cones) are radially oriented and accommodated on the retinal pigmented epithelium. Cone photoreceptors are responsible for high-acuity and color vision, while rod photoreceptors allow us to see at low light intensity levels. Photoreceptors communicate to horizontal and bipolar cells in the inner nuclear layer, which signal to amacrine cells and finally to retinal ganglion cells (RGCs). The axon of each RGC extends from the location of the cell body to the optic disc, where they are collected and directed out of the eyeball to form the optic nerve. At the optic chiasm, the two optic nerves are split roughly in half

and recollected to have the same field of view side bundle together. The axons finally arrive at the lateral geniculate nucleus of the thalamus, where the visual input is sent to the occipital lobe at the primary visual cortex (V1) for information integration and generation of visual experiences.

Visual acuity is an important term when talking about the visual system. Visual acuity refers to the ability of the eye to distinguish shapes and the details of objects at a given distance. Typically, a 20/20 visual acuity value is considered perfect vision. High visual acuity vision is found in the fovea (central point on the retina), which is densely saturated with cone photoreceptors. It is very important for reading and faces recognition [11]. Nevertheless, peripheral vision is characterized by a lower concentration of cone photoreceptors compared to the fovea. Although peripheral vision has a lower resolution, it is just as much valuable as the macular high visual acuity. In fact, humans can get information about the location of objects, their overall shape and size, their color, and their movements from a field of view much wider than 10° . The extent of a normal field of view is up to about 200° horizontally (with 120° binocular) and 130° vertically. Vision is necessary up to the far periphery, even if with poor resolution, in order for us to move in the environment, avoid obstacles, and react to incoming dangers.

1.2 Vision impairments

According to the World Health Organization, 285 million people are estimated to be visually impaired worldwide [12]. Of these, 39 million are blind and 246 million have low vision. In various parts of the world, legal blindness is defined as visual acuity of 20/200 or field of view no greater than 20° in the better eye with best correction possible.

Vision disorders are a major health issue for society and for the individual's quality of life. Low vision is uncorrectable vision loss that interferes with the possibility to perform the basic self-care activities of daily, sustain independency, and experience everyday-life visual emotions (like seeing someone's face and expressions). The leading causes of blindness are primarily age-related eye diseases such as age-related macular degeneration (AMD), cataract, retinitis pigmentosa (RP) and glaucoma [13–15].

Photoreceptors are particularly susceptible to cellular stress and loss of photoreceptors due to degeneration is a major cause of vision loss, resulting in dysfunctional light detection, transduction, and transmission. These inherited disorders can affect either rods or cones

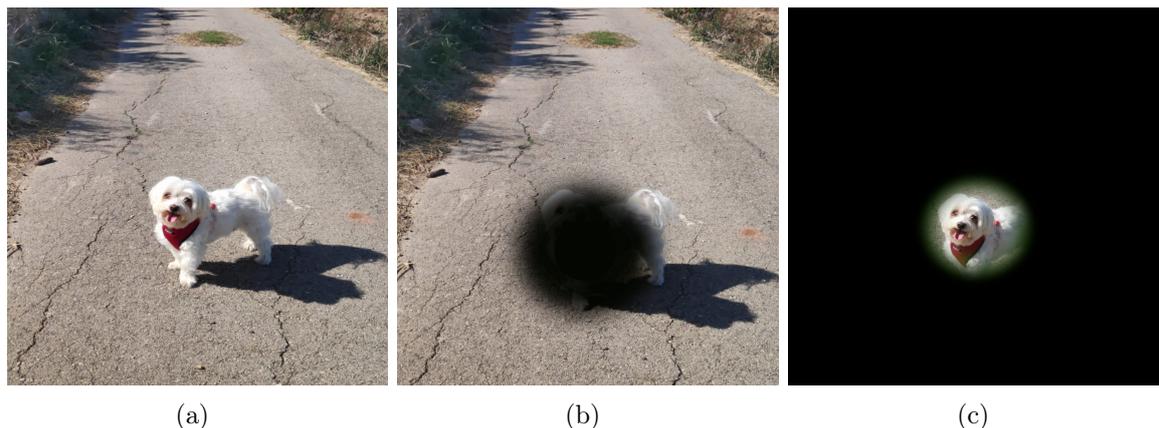


Figure 1.2: Visual impairments reflected in a beautiful dog image. (a) Normal vision. (b) Age-related macular degeneration (AMD). (c) Retinitis pigmentosa (RP).

primarily, as in RP. The most common form of RP is a rod–cone dystrophy, affecting night blindness, followed by progressive loss in the peripheral field of view in daylight, eventually leading to blindness after several decades. Another important degenerative disease of the retina is the AMD. This retinal disease is characterized by sudden acuity loss resulting from untreatable submacular neovascularisation. Therefore, the outer retinal degeneration in AMD and RP leads to the appearance of stains in the macular or peripheral regions of the retina, respectively. Consequently, the image perceived by the patients is affected in the center for AMD or on the sides for RP, which develops the so-called “tunnel vision” before complete blindness, as can be seen in Figure 1.2.

1.3 Sight restoration technologies

At the moment, there is no cure for blindness and conditions like AMD and RP but to prevent or mitigate the disease advancement. Researchers are been trying to come up with ways to restore sight in these individuals. One way would be to use gene therapy [16, 17] to simply replace the faulty genes with functioning ones. Another way would be to induce small optogenetic proteins [18, 19] that make the surviving cells sensitive to light again. The third idea is based on an electrical prostheses [20, 21] where directly elicit responses in surviving retinal cells, analogous to cochlear implants.

Although the development of efficient biomedical and engineering concepts is promisingly advancing, for some of the emerging treatments the path for clinical and commercial applications remains delayed. However, visual prostheses, more precisely retinal prostheses,

have finally received regulatory market approval in Europe and United States, becoming a valuable option to artificially induce visual perceptions in blind patients. A visual prosthesis is an visual device intended to restore functional vision in people suffering from partial or total blindness. A number of researchers are studying the partial restoration of sight to blind people through the electrical stimulation of a component of the visual system. The idea of electrically stimulating the human visual system was first described by Franklin [22] and Leroy et al. [23]. Leroy et al. [23] studied visual sensations of light by passing an electrical charge through the eye of a blind man. Later, in 1929, it was shown that stimulating the human visual cortex led to the perception of spots of light referred to as “phosphenes” (see Figure 1.3). The phosphenes created by prosthetic vision can have various brightness, shape, size, and blurring depending on the stimulating electrodes (size, shape, pitch, and field lines distribution), stimulation parameters (current pulse intensity, length, and frequency), and anatomical target. Since then, the idea of restoring sight to the blind is closer.

1.3.1 Principles of prosthetic vision

A visual prosthesis system needs to perform a specific sequence of actions in order to simulate the complex series of events that take place between the outside world and the visual cortex, such as image capture, image processing, microelectrode array stimulation, which in turn must deliver a suitable stimulation current (directly or indirectly) to the RGC (see Figure 1.3).

Image capture

The image capture component is based on capturing information from the visual scene using an external video camera, which is usually mounted on glasses. The video camera then transmits data directly to an external video-processing unit (VPU). However, due to the fixed camera position, there is a risk of discrepancy between its orientation and that of the eye, risking inaccuracies in the perceived spatial location of an object of interest following retinal stimulation. Various solutions to this problem include both the incorporation of an eye tracker to coordinate the direction of gaze with the stimulation pattern, or placement of the camera system intraocularly, for which there have been some preliminary attempts to design such a system [24].

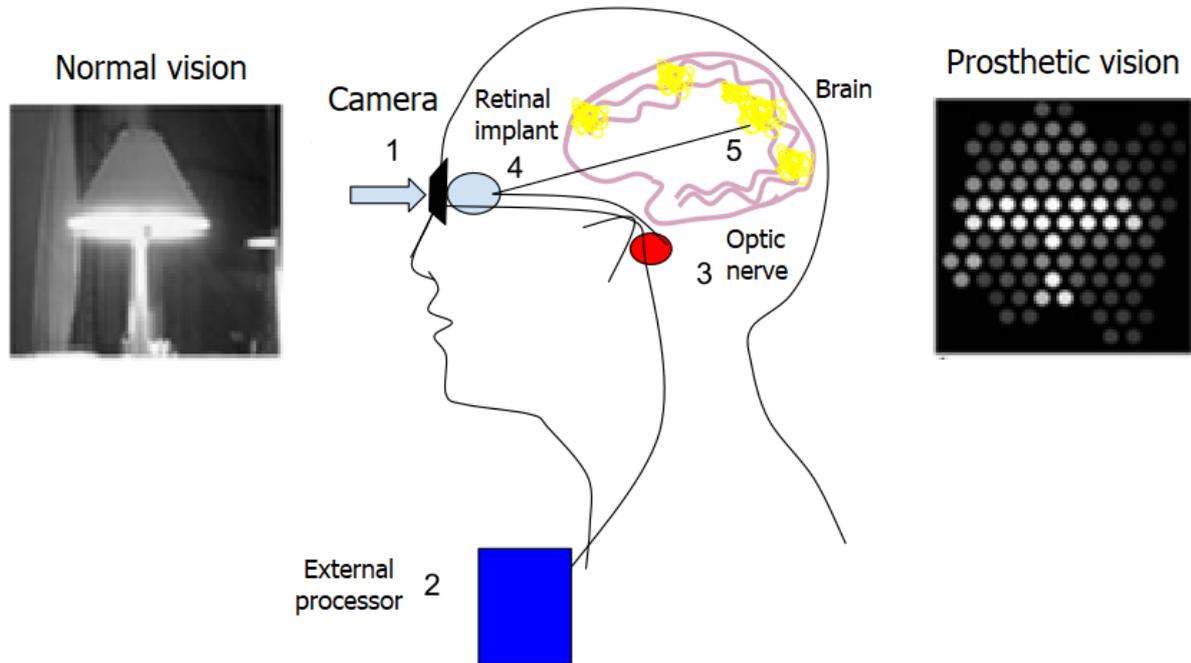


Figure 1.3: Prosthetic vision. Initially, the patient receives an implant in their retina. Glasses worn have a miniature camera which then send signals to the implanted chip located in the retina. Then, the chip converts these signals into electrical impulses that can then be sent to the optic nerve and processed as an image in the brain.

Image processing

The image processing takes place within the external VPU, which receives a high-quality signal from a camera system and transforms the data into a set of commands to be wirelessly transmitted to the microelectrode array to generate a particular stimulation pattern. In systems such as the Argus II retinal prosthesis, there are only 60 microelectrodes [25]. This suggests that even with a perfect contact of the electrodes and an accurate perception of the retinotopic phosphene, the resolution of the image would be still be low. To deal with this, several researchers have emphasized the development of software algorithms that can filter the most relevant parts of a visual scene, before creating simple stimulation configurations that map the object of interest onto the array [26–28].

Data transmission

Normally, the delivery of visual information that is sent to the microelectrode array is done wirelessly (via inductively coupled coils). An AC current passing through the external coil induces an AC voltage in the internal coil, which can subsequently be converted

into DC power. A capacitor in series with the secondary coil permits amplification of the received voltage by creating a tuned resonance at the transmitter frequency, thus supporting efficient power transfer while minimizing the body's exposure to radiation. With this model, the data transfer capacity is sufficient to support the resolution and refresh rate of today's systems [29].

Microelectrode array stimulation

Electrode material, size, shape, spacing, tissue contact and the anatomical position of the array are considerations to take into account when designing an electrode array to deliver electrical currents to RGCs in order to replicate the spatial resolution of the natural retina. Electrode shape may significantly affect the integration of the device by creating an environment where retinal tissue can migrate around the array and create a close interface. Size, shape and contact of the electrode at the tissue interface affect the charge density per unit area. It is known that as the size of the electrode becomes smaller, the concentration of the current increases exponentially, which can result in target tissue damage.

Regarding to the electrode position, there are three main types of retinal implants by placement because of the easy ocular access: epiretinal (on the retinal surface, held with retinal tacks), subretinal (between the retinal pigment epithelium and neural retina, replacing photoreceptors), and suprachoroidal (between the sclera and the choroid). Suprachoroidal implants have the disadvantage that they are placed relatively far from the target cells. Even so, they have shown the appearance of phosphenes in clinical studies and have a lower risk in terms of retinal damage and detachment. In spite of that, the two most common retinal implant placements are epi- and subretinal. However, the main disadvantages of subretinal implants include impaired nourishment of the inner retina due to the creation of a mechanical barrier between the outer retina and the choroid and trauma to the retina during implantation. These prostheses also show poor dissipation of heat and therefore could damage the retina. In contrast, among the advantages of epiretinal prostheses are the possibility of wide coverage of the field of view (allow the restoration of vision in both central and peripheral fields), easy placement and better heat dissipation, compared to the great disadvantage of subretinal implants of presenting a high risk of retinal detachment. Another advantage of epiretinal implants is that a

camera can process signals before they reach the implant. This allows to optimize the signal quality and the use of artificial intelligence techniques, which may lead to improved visual perception.

1.3.2 Retinal prostheses

In the last decade, several groups have developed retinal prostheses (see Figure 1.4). Only four devices have been granted CE Mark for commercial use in the European Economic Area: Retina Implant Alpha IMS (first-generation device, Retina Implant AG, Reutlingen, Germany) [30, 31], Retina Implant Alpha AMS (second-generation device, Retina Implant AG) [32, 33], IRIS II (Pixium Vision, Paris, France) [34] and Argus II Retinal Prosthesis System (Second Sight Medical Products Inc, Sylmar, CA), which was the first visual prosthesis to become commercially available: it received the European conformity mark in 2011 and FDA approval was granted in early 2013 for humanitarian use in the USA [25, 32].

Alpha IMS and AMS Subretinal implant

The Alpha IMS and Alpha AMS devices both consist of a microchip on a polyimide foil and a cable for power supply and signal control (see Figure 1.4(a)). The microchip consists of 1500 (Alpha IMS) or 1600 (Alpha AMS) independent photodiode-amplifier-electrode units. Each electrode transforms the local luminance information into an electrical current that is amplified and transferred to the adjacent bipolar cells via an electrode. The light that falls on the chip is used to directly control the stimulation current amplitude in each pixel. One of the advantages of this system is that a large number of electrodes can be addressed, no external camera is needed, and natural eye movements can be used by the patients to locate and fixate on objects [35].

IRIS II Epiretinal implant

IRIS II incorporates innovative and distinctive features: A bio-inspired camera intended to mimic the functioning of the human eye by continuously capturing the changes in a visual scene with event-based camera with asynchronous pixels, and unlike an imaging sensor that takes a sequence of video frames with largely redundant information (see Figure 1.4(b)). It is composed by an epiretinal implant with 150 electrodes secured on the

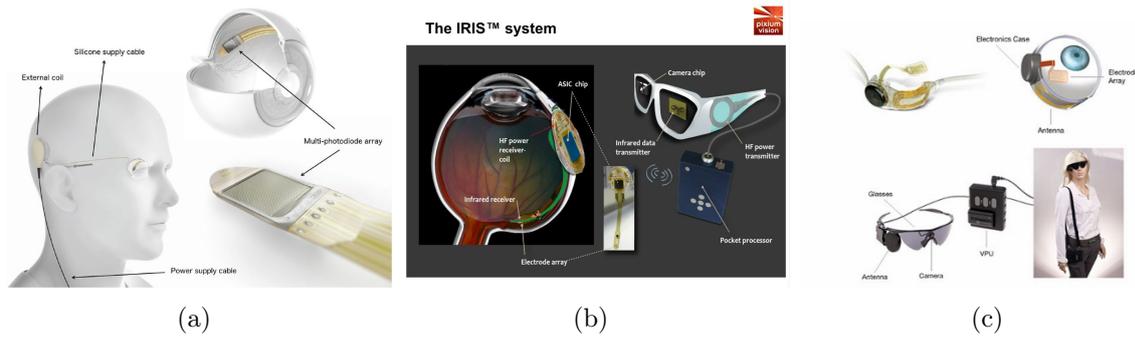


Figure 1.4: Retinal prostheses. (a) Alpha AMS Subretinal implant (Alpha AMS; Germany, CE marking 2016), (b) IRIS from Pixium Vision (France, CE marking 2016) and (c) Argus II from Second Sight Medial Productst (USA, CE marking 2011, FDA-PMA 2014).

retinal surface by a patented support system that is intended to allow future replacements or upgrades.

Argus II Epiretinal implant

The Argus II device consists of three internal components and three external components (see Figure 1.4(c)). The internal components implanted around the eye include a receiver coil, electronics and an electrode array which has 60 platinum arranged in a 6×10 grid (diameter = $200 \mu m$) spaced $575 \mu m$ (center-to-center). The array of electrodes covers about 20° of field of view (diagonally). The electrode array, which is made of a material that allows a perfect adjustment to the curvature of the retina, is attached to the retina over the macula with a retinal tack.

The external components consist of a video camera that is mounted on glasses allowing real-time image capture, a VPU that is placed on the patient's belt and a coil on the side arm of the glasses to transmit data between the internal and external components using radio frequency. The patient is wearing a pair of glasses with a small camera embedded. The VPU has adjustable settings that transform the images from the camera into electrical stimulation, which is then transmitted using an antenna via the coil on the side of the glasses to the receiving coil in the sclera. This information is then sent via the cable to the electrode array on the patient's retinal surface stimulating remaining viable inner retinal cells. This artificial stimulation makes its way via the optic nerve and lateral geniculate nucleus through the visual system to the occipital cortex and the induced vision is perceived as light patterns in the visual cortex [36].

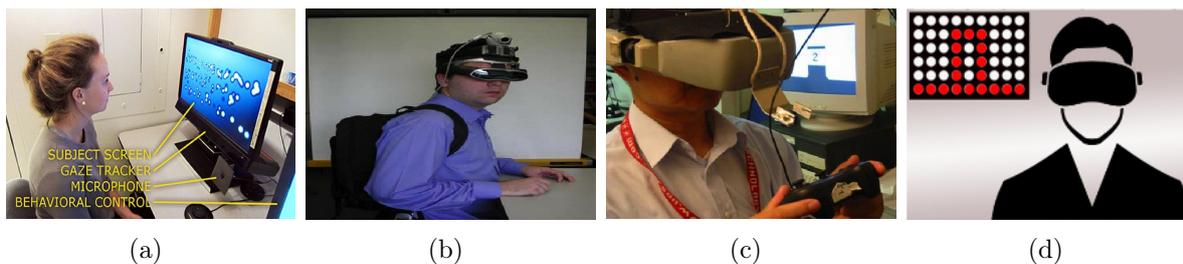


Figure 1.5: Simulated Prosthetic Vision (SPV) setups. (a) The stimulus is displayed in a computer screen in front of the subject [37]. (b) Subjects carry a PC video camera placed in a modified adjustable head strap in a midline position above the eyes. The laptop is placed in a specially designed computer backpack, making the system completely portable [38]. (c) Subjects view pixelized imagery in a head-mounted display [39] or (d) through a virtual patient [40].

1.4 Simulated Prosthetic Vision

Many clinical trials related to visual prosthesis are carried out over long periods of time and in very closed protocols. Therefore, implanted subjects are not available to perform other experiments and test other possible versions of implants. However, it is possible to use Simulated Prosthetic Vision (SPV).

The simulation is based on the subjects' descriptions of the phosphenes they have perceived. The SPV setup contains the characteristics of the implant to simulate, as well as the shape of the desired phosphenes. The function of a SPV system is to show a representation that corresponds as closely as possible to the implant to be simulated. Different SPV setups have been used in the literature [37, 38, 40] (see Figure 1.5). One option is to display the visual stimulus on a computer screen in front of the subject [37]. Another option is to carry a PC video camera placed in a modified adjustable head strap in a midline position above the eyes. The laptop is placed in a specially designed computer backpack, making the system completely portable [38] or subjects view pixelized imagery in a head-mounted display [39]. The prosthetic vision models currently used simulate the shape, intensity, size and regularity of phosphene to simplify the possible range of perceptions of volunteers with normal vision and, therefore, represent the best scenario for a blind patient in these terms.

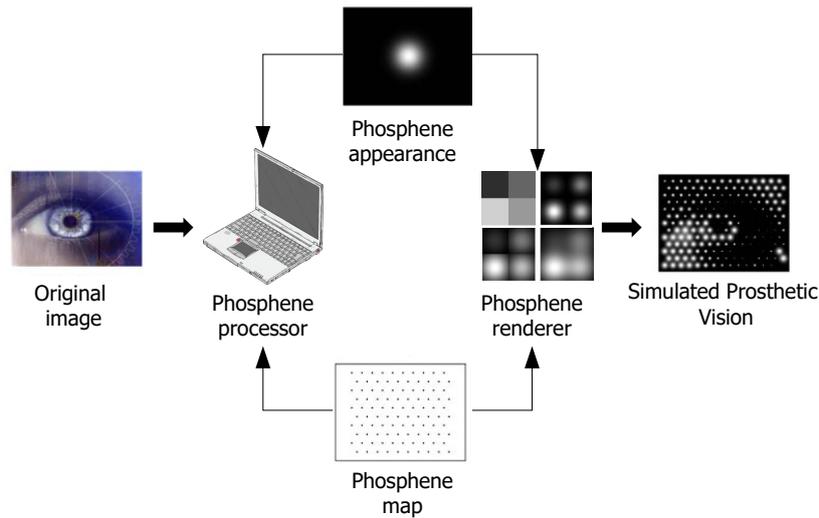


Figure 1.6: Visual models of phosphenes. First, the phosphene processor filter the original image and phosphenized based on the selected phosphene appearance and phosphene map. The output of the phosphene processor is visualized at the phosphene renderer and produce the final simulated prosthetic vision.

1.4.1 Visual models of phosphenes

Prosthetic vision is built upon phosphenes. A phosphene can be described as ‘a visual sensation caused by means other than stimulation of the visual system by light’. It can be directly induced by mechanical, electrical, or magnetic stimulation of the retina or visual cortex, or by random firing of cells in the visual system. But despite the increasing clinical and commercial use of these devices, the perception of implant users remains poorly understood. The appearance of individual phosphenes is highly variable not only across subjects but also across electrodes within a subject, with subjects typically reporting seeing distorted and often elongated geometric shapes that fade quickly over time [41–43].

Taking this idea of phosphenes, many investigators have implemented simulations (or visual models) of the anticipated form of restored vision [43–45]. More interesting, Beyeler et al. [44] have developed “pulse2percept”, an open-source Python implementation of a computational model that predicts the perceptual experience of retinal prosthesis patients across a wide range of implant configurations and phosphene models. Usually, a simulated phosphene can be divided into four modules: Phosphene appearance, phosphene map, phosphene processor and phosphene renderer (see Figure 1.6).

Phosphene appearance

A phosphene appearance is a set of phosphene visual profiles with different sizes, luminance, color, and other visual attributes [45]. The common form of phosphenes is a small, round, colored spot of light in the field of view.

Gaussian luminous distribution is often used to model the single phosphene through feedback from prosthesis wearers. The distribution function of the 2-D Gaussian profile presents a luminance profile where the luminance is the brightest at the center and gradually decays to the periphery. Up to 12 levels of luminance have been observed in the literature [46]. Typically, eight different luminance levels are used in SPV according to the number of luminance levels attainable in most human trials using retinal prostheses [45, 47].

Phosphene map

The phosphene map locates phosphenes in the field of view based on information provided by the implant recipient. Despite well-known regular mappings between the stimulation sites and the field of view such as the retinotopic field of view organization, the phosphene maps from regular lattices of stimulating electrodes are often quite distorted. However, the regular lattices adopted by SPV investigators are commonly square or hexagonal.

Phosphene processor

The phosphene processor is a simple image processing routine taking the gray level values at phosphene locations on the visual scene corresponding to the phosphene map to obtain the size and luminance for each phosphene. A simple extension of this used in SPV is a mean filter, which converts the input image directly to the phosphene map by averaging the brightness on the region covered by each phosphene. On the other hand, with a Gaussian filter, the phosphene will be weighted with the luminance information towards the center of the receptive field.

Phosphene renderer

The phosphene renderer composes a phosphene field with the phosphenes retrieved by the phosphene processor and determines how realistic the simulation is compared to

implantee reports. This step is a type of visualization module responsible for painting the SPV on the screen. Some investigators simply rendered phosphenes as square pixels immediately adjacent to each other. This simulation is very unrealistic compared to the actual prosthetic vision. In addition, additional perceptual artifacts in the square corner and edges where grayscale levels change gradually will affect the quality of perception and performance using SPV. A more realistic simulation is to separate each phosphene with an empty space in the presentation to simulate gaps between phosphenes, as most investigators have done.

Dropout

Some amount of electrodes cannot elicit phosphene successfully due to the degenerating ganglion cells in retina of the implant recipients, and some might not be able to elicit phosphenes after long-term implantation, which account for the dropout phenomenon of phosphene map. Normally the dropout effect is simulated as a random percentage of the corresponding amount of dotted phosphenes from the phosphene map in SPV being turned off.

1.4.2 Experimentation with SPV

SPV has been used as a tool for more than two decades to assess the characteristics of implants necessary to enable future people with implantation to perform different tasks in their daily lives. The most experimented tasks in SPV are reading, face recognition, object recognition and localization and navigation (see Figure 1.7).

Reading

Reading is one of the most studied tasks in SPV, reflecting its importance for implanted people. So far, implanted people had previously acquired blindness; therefore, they could read before losing their sight. Some reading studies have focused on character recognition [51–53]. Others studied focused on how many electrodes were needed to limit the number of errors in individual words, concluding around 300 electrodes [54], or 600 electrodes to allow paragraph or full text reading [55, 56]. For low resolution implants, it is possible to read at a speed of 15 words per minute with only about 40 phosphenes [57].

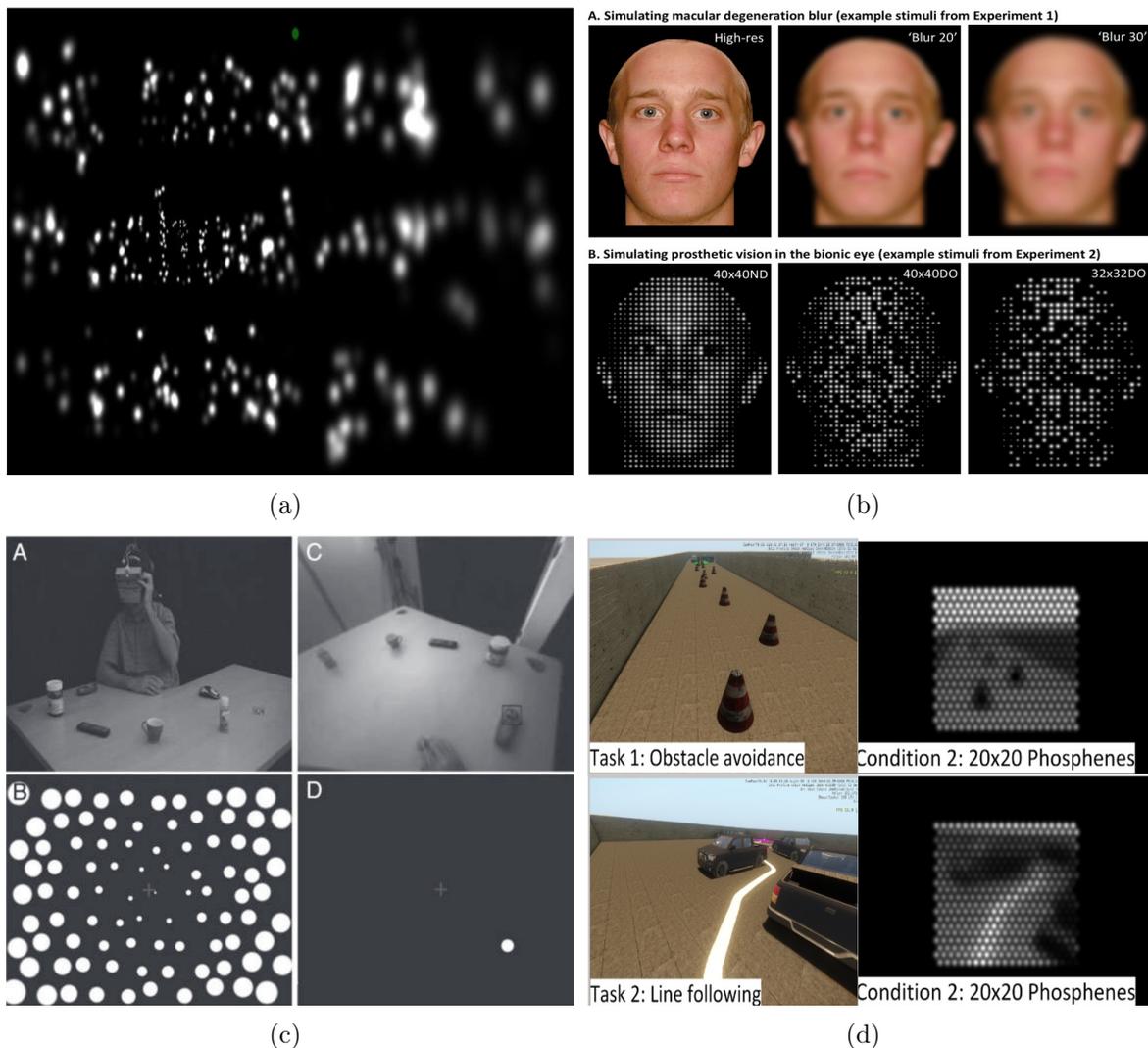


Figure 1.7: SPV approaches for different tasks. (a) Reading [37], (b) face recognition [48], (c) object localization [49] and (d) navigation [50].

Face Recognition

Another very explored area in SPV concerns the face recognition and localization. Facial expressions are very important for social interactions [58], communication containing a large part of non-verbal communication [59]. Restoring the perception of faces is therefore essential for the blind. The first SPV experiment involving faces focused on discrimination between faces, studying the minimum resolution necessary for the recognition of faces [60, 61]. According to this study, difficulties in recognizing faces begin when the electrode array contains less than 100 electrodes. Other experiments focus on the use of image processing algorithms to modify the contrasts of a face so that it can be recognized more easily [62–64]. For instance, Chang et al. [62] generated a contrast-enhanced image

and an edge image by applying the Sobel edge detector. They found that the subjects recognized a distinctive face especially more accurately and faster when using the proposed contrast and edge-enhancement method than the other given facial images even under low-resolution prosthetic vision.

Object Recognition and localization

Complex tasks such as object recognition and localization are still very limited with visual prostheses. Denis et al. [65] tested the object recognition and localization in the surrounding only highlighting a few representative phosphenes. Subjects were asked to search for a specific object placed among nine others on a table. Their results suggest that the efficiency of current electrode arrays can be improved by including an algorithm for locating contextual objects in the visual prosthesis. Three years later, Macé et al. [49] proposed an alternative method based on object recognition in order to display only relevant information. They evaluated this approach in a reach-and-touch task requiring object discrimination and localization using a real-time object recognition system. When an object of interest was recognized and localized within the camera image, its position was displayed by switching on a unique phosphene at the corresponding location. Only nine electrodes were necessary for object localization.

Image processing has also been proposed for object recognition tasks. For example, Han et al. [66] proposed a foreground extraction method based on a saliency model to obtain a foreground object. Their results showed that the foreground extraction method had prominent advantages in comparison with direct pixelization in terms of recognition accuracy and efficiency. Li et al. [67] applied an image processing strategy for optimizing the visual perception focusing on recognition of the object of interest. They show how using a saliency segmentation method and image processing strategies can automatically extract and enhance foreground objects, and significantly improve object recognition performance. In more recent work, Han et al. [68] combine deep learning-based scene simplification strategies with a psychophysically validated computational model of the retina to generate realistic predictions of simulated prosthetic vision and measure its ability to support scene understanding of sighted subjects. (virtual patients) in a variety of outdoor settings. Their findings show that object segmentation can better support scene understanding than models based on visual prominence and monocular depth estimation.

Navigation

Navigation is an important component of the self-reliance of the blind, and SPV studies have focused on this issue, more specifically on obstacle detection and avoidance.

Dagnelie et al. [38] performed two experiments based on navigation using prosthetic vision. In the first part of the experiment subjects traveled similar routes around a suite of offices with simulated implants of 4×4 , 6×10 and 16×16 dots. In the second experiment, they used ten different virtual buildings adding dynamic noise and removed a subset of phosphenes from a 6×10 array. Their findings suggest that a retinal implant with as few as 60 electrodes may provide independent orientation skills to the blind. Wang et al. [39] assessed virtual maze navigation performance with SPV in gaze-locked viewing, under the conditions of varying luminance contrast, background noise, and phosphene dropout. The results suggest that the simulated gaze-locked can be helpful for wayfinding in simple mobility tasks, though phosphene dropout may interfere with performance. Van Rheede et al. performed a similar study on navigation [69]. They developed a simulation paradigm that used a head-mounted camera and eye tracker to lock the simulation to the point of fixation. They evaluated visual acuity, object recognition and manipulation, and wayfinding under SPV in three ways of optimizing the information varying the field of view.

Other researchers have focused on techniques based on depth visualization, more concretely, encoding distance as a function of luminance. The idea is to represent the environment by translating the distance to an element of the visual scene into light intensity. So, a phosphene no longer represents the luminosity of a certain area, but the average distance of this area [70–72]. This strategy allowed subjects to have better mobility performance than with conventional rendering when obstacles had to be avoided. Based on these results, McCarthy et al. [73] proposed to use ground surface segmentation to enhance the perception of obstacles in low to medium resolution prosthetic visual representations. In 2012, McCarthy et al. [74] proposed another rendering to try to overcome renderings based on the luminance of objects or their distance in the image. The luminance of the phosphenes indicated the time remaining before the object made contact with the subject. Another augmented rendering has been proposed by Parikh et al. [75]. The proposed rendering indicated the salient objects present in the field of view of the camera. To this field of view representing a salience map, the authors proposed

adding a clue indicating the direction of the nearest object. More recently, Zapf et al. [50] proposed to use a white line on the ground in order to pass between stationary cars.

1.5 Motivation and challenges

Millions of people worldwide could potentially benefit from more efficient visual prostheses. With current prostheses, reasonable expectations in vision restoration are limited to large object recognition, localization, and movement detection, large letters reading, and large obstacles avoidance. Furthermore, extremely long times are necessary for the patients to detect, analyze, and react to the scene in front of them, mostly because of the poor prosthetic resolution and limited field of view. Although these results are still very promising and mark an important milestone artificial vision, with a low resolution implant, rendering the visual scene with no specific processing leads to visual overcrowding.

The perception system used by most of these visual prostheses is based on the acquisition of images with an external camera. This configuration (camera+prosthesis) allows exploring more advanced computer vision and deep learning techniques to enhance the semantics and the relevance of the information displayed to the patient.

Our motivation is that powered visual prostheses could react in synchronicity with their users by recognizing and anticipating the environment in front of them.

1.6 Goals and Contributions

Once we have asserted our motivation of the thesis, it is time to define the specific goals to pursue, and the extent of the work developed in these four years in those lines of research. In particular, the thesis can be divided in five main lines of work.

- **Schematic representation of indoor scenes:** The ability of object recognition in real-life scenarios is severely restricted for prosthetic users due to the low resolution, limited field of view, and the low dynamic range of the visual perception. This results in huge loss of information occurred when presenting daily scenes. To overcome the limitations, we aim to optimize the visual information in the SPV. We want to build a schematic representation of indoor environments for simulated phosphene images. To do that, we will use a variety of convolutional neural networks for extracting and

conveying relevant information about the scene such as structural informative edges of the environment and silhouettes of segmented objects.

- **Influence of field of view:** The field of view in prosthetic vision is linked to the retinal coverage of the prosthesis, while the visual acuity to the density of the stimulating electrodes. Therefore, the number, density, and coverage of the electrodes is an important design parameter for retinal prostheses to provide adequate artificial vision in daily life. We aim to investigate the influence of field of view with respect to spatial resolution in visual prostheses. We took advantage of the proposed virtual-reality system based on a head-mounted display and panoramic scene to monitor the performance of healthy participants under SPV with variable pixel density and field of view angle. This system acts as an electronic visual aid that attaches to the user's head and presents information directly to the user's eyes. Our system allows replications and extensions of experiments in a simple way.
- **Augmented reality navigation:** Implanted patients still require constant assistance for navigating from one location to another. Hence there is a need for a system that is able to assist them safely during their journey. However, the development of navigation devices capable of guiding the blind through different scenarios has remained a challenge. We aim to investigate a new augmented navigation system with obstacle avoidance for guidance in visual prosthesis. By using a route planning algorithm, the system guides the subject through a shorter, obstacle-free path.
- **Visual acuity assessment with VR:** The visual acuity of prosthetic vision is limited by various factors from both engineering and physiological perspectives. One of the main causes of low visual acuity is the limited spatial resolution that can be achieved by electrical stimulation with existing retinal implants. The size of the electrodes in today's retinal implants is often much larger than the size of the neurons in the retina, and the number of electrodes is low. We aim to investigate the influence of field of view with respect to spatial resolution in visual prostheses measuring visual acuity. We implemented prosthetic vision in a virtual reality environment in order to simulate the real-life experience of using a retinal prosthesis.
- **Spiking neural network:** While using artificial neural network algorithms may be sufficient to improve perceptions with visual prostheses, employing better neural

coding algorithms would improve the performance of retinal prostheses. The reconstruction of visual scenes can be significantly improved by adding an encoder that converts the input images into the spiking codes used by retinal ganglion cells and use these codes to stimulate electrodes. We aim to develop a spiking neural network model for visual coding to obtain better computational algorithms and improve the performance of future retinal prostheses.

Next we present this thesis' contributions leading towards the aforementioned goals.

1.6.1 Schematic representation of indoor scenes

The goal of this part of the thesis is to build a schematic representation of indoor environments for simulated phosphene images since the phosphenic images produced by the current implants have very limited information bandwidth due to the poor resolution and lack of color or contrast. Thus, the ability of object recognition and scene understanding in real environments is severely restricted for prosthetic users.

We propose a new visual representation of indoor environments for prosthetic vision, which emphasizes the scene structure and object shapes [76, 77]. The proposed method combines a variety of convolutional neural networks for extracting and conveying relevant information about the scene such as structural informative edges of the environment and silhouettes of segmented objects. Our results demonstrate that our method is well suited for indoor scene understanding over traditional image processing methods used in visual prostheses. The key idea of our current results is that, with only a few significant elements of the scene, it is possible to obtain a good perception of the environment, even in complex and occluded scenes. These methods and their evaluation are presented in detail in Chapter 2. In the benefit of the visual prostheses design community and for reproducibility purposes, we have created an image Dataset of phosphene images which can be found in the dataset available online: <https://doi.org/10.6084/m9.figshare.11493249.v4>.

1.6.2 Influence of field of view

The second part of the thesis addresses the problem of how to improve the quality of electrode arrays, in terms of features, such as response time, resolution or size. To the best of our knowledge, this is what the latest developments in retinal implant technology

are investigating. Understanding the influence of these parameters in the perception results can guide prostheses research and design. The general objective is to address the question of how to evaluate and predict the utility and functionality in terms of patient benefit with respect to design parameters.

We evaluate the influence of field of view with respect to spatial resolution in visual prostheses. To meet this goal, we measure the accuracy and response time in a search and recognition task [78]. To validate this approach, twenty-four normal participants were asked to find and recognize usual objects, such as furniture and home appliance in indoor scenes. For the experiment, we use a new simulated prosthetic vision system that allows simple and effective experimentation. Our system uses a virtual-reality environment based on panoramic scenes. The simulator employs a head-mounted display which allows users to feel immersed in the scene by perceiving the entire scene all around. Our experiments use public image datasets and a commercial head-mounted display. We have also released the virtual-reality software for replicating and extending the experimentation <http://webdiis.unizar.es/~rmcantin/index.php/Research/Vrfov>. This method and its evaluation is presented in detail in Chapter 3.

1.6.3 Augmented reality navigation

The goal of this third part of the thesis is to develop a new augmented reality navigation system with obstacle avoidance for guidance in visual prosthesis since unfortunately implanted patients still require constant assistance for navigating from one location to another. Hence there is a need for a system that is able to assist them safely during their journey.

We propose an augmented reality navigation system for visual prosthesis that incorporates a software of reactive navigation and path planning which guides the subject through convenient, obstacle-free route. It consists on four steps: locating the subject on a map, planning the subject trajectory, showing it to the subject and re-planning without obstacles. We have also designed a simulated prosthetic vision environment which allows us to systematically study navigation performance. Twelve subjects participated in the experiment. Subjects were guided by the augmented reality navigation system and their instruction was to navigate through different environments until they reached two goals, cross the door and find an object (bin), as fast and accurately as possible. Results show

how our autonomous navigation system help navigation performance by reducing the time and distance to reach the goals, even significantly reducing the number of obstacles collisions, compared to other baseline methods.

1.6.4 Visual acuity assessment with VR

The goal of this fourth part of the thesis is to study the optimal number of the electrodes and field of view that a visual prosthesis should have to provide adequate visual acuity in daily activities. This is still an open question and an important design parameter needed to develop better implants.

We asses visual acuity in visual prostheses by taking advantage of a virtual-reality system in order to simulate real-life experience of wearing a visual prosthesis. Subjects were required to identify computer-generated Landolt-C optotypes and different stimulus based on luminance, time-resolution of luminance, localization of light and motion commonly used for visual acuity examination in the sighted. Ten normally sighted participants volunteered for the study. Performance of subjects in correct identification and reaction time of Landolt-C gap orientation and stimulus is reported. Visual acuity was estimated by fitting the performance versus Landolt-C gap size. The results of our study showed that in all the tasks the field of view played the most significant role in improving the performance of the subjects. The optimal visual acuity was 1.43 logMAR and was obtained for the condition of 20° of field of view and 1000 resolution. The design of new retinal prostheses should take into account the relevance of the restored field of view to provide a helpful and valuable visual aid to profoundly or totally blind patients.

1.6.5 Spiking neural network

The goal of the last part of the thesis is to develop a spiking neural network model for visual coding to obtain better computational algorithms and improve the performance of future retinal prostheses.

We present a spiking neural network which relies on a combination of biologically plausible mechanisms and uses unsupervised learning scheme, i.e., the weights of the network learn patterns of the input images without using labels. Our model is simple, fast, energy-efficient and bio-inspired and uses spike-timing-dependent-plasticity and lateral inhibition with one trainable layer. We use a pre-processing scheme based on feature

extraction using multiple spatial frequencies in the input sample and we combine them at the output of the spiking neural network for a classification task.

1.7 Organization of the Document

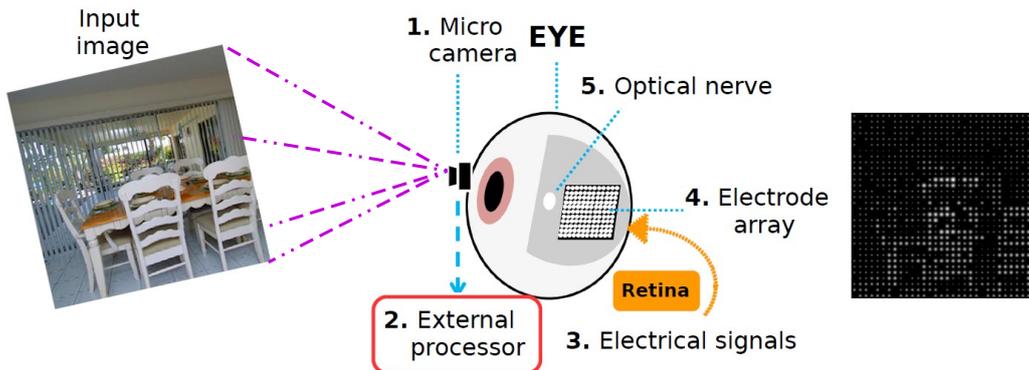
The following five chapters relate the research conducted in the direction of the main contributions of this thesis regarding schematic representation of indoor scenes (Chapter 2), influence of field of view (Chapter 3), augmented reality navigation (Chapter 4), visual acuity assessment (Chapter 5) and spiking neural network (Chapter 6) with some preliminary results. Each of these chapters intends to be self-explanatory and contains its specific related work as well as its respective conclusions and discussions. In Chapter 7 we seek to outline the general conclusions of the thesis, as well as our vision of its future directions. A summary of the results can be found in Chapter 8.

Life in a new light

Chapter 2

2 Schematic representation of indoor scenes

Prosthetic vision is being applied to partially recover the retinal stimulation of visually impaired people. However, the phosphenic images produced by the implants have very limited information bandwidth due to the poor resolution and lack of color or contrast. The ability of object recognition and scene understanding in real environments is severely restricted for prosthetic users. Computer vision can play a key role to overcome the limitations and to optimize the visual information in the prosthetic vision, improving the amount of information that is presented. We present a new approach to build a schematic representation of indoor environments for simulated phosphene images. The proposed method combines a variety of convolutional neural networks for extracting and conveying relevant information about the scene such as structural informative edges of the environment and silhouettes of segmented objects. Experiments were conducted with normal sighted subjects with a Simulated Prosthetic Vision system.



2.1 Introduction

Retinal degenerative diseases such as retinitis pigmentosa and age-related macular degeneration cause loss of vision due to the gradual degeneration of the sensory cells in the retina [79, 80]. Retinal prostheses are currently the most promising technology to improve vision in patients with such advanced degenerative diseases [25, 81–83]. These devices elicit visual perception by electrically stimulating retina cells. As a result, implanted patients are able to see patterns of spots of light called *phosphenes* that the brain interprets as a visual information [45, 84, 85]. Current retinal prosthetic devices are limited to hundreds of electrical receptors, which produce a very limited visual elicitation [86–88]. From the actual technologies for retinal implants [89], one of the most active line of research is based on implants with a micro camera that captures external stimuli and a processor that converts the visual information in microstimulations in the implant. Following the computer image paradigm, we can say that the visual information evoked by the implants has very low spatial resolution and very limited dynamic range (only few levels of stimulus intensity are perceived as different) [90–92]. Intuitively, from an information theory perspective, the process from the external sensor input to the implant stimuli is analogous to taking a high definition image and convert it to a low resolution, grayscale image with just a few grey levels. Thus, a large amount of visual information is lost. Prosthetic vision allows users to recognize objects with simple shapes, to see people’s silhouettes in bright light or detect motion [93], but high level tasks require more precise visual cues and a deeper interpretation of the information.

Recent developments in implants might result in an improved resolution and performance of the visual elicitation [94], but quality would still be several orders of magnitude lower than a current micro camera. Alternatively, the visual information gathered by the external camera could be processed prior to being transferred to the retinal electrodes. Image processing can be used to extract and highlight relevant information from the external camera. This information can be presented with visual cues that help to understand the perceived scene by the implanted subject. Several studies have already been conducted testing specific cues for object recognition [47, 49, 67, 69, 95–97], reading [37, 54, 55, 98], facial recognition [48, 64] or navigation [38, 75, 99–101] in the context of prosthetic vision.

One of the most basic image processing tasks from the cognitive, but also from the

computational level, is the segmentation of the image in different regions [102–104]. From a statistical point of view, this corresponds to the problem of *clustering*. Rooted on the Aristotelian laws of association, early research in perception from *Gestalt* psychologists found the importance of the *principles of grouping* [105]. These principles state that our brain tends to group image elements based on proximity, color, shape or other similarities. Although some of the Gestalt ideas are controverted, the principles of grouping have been supported by posterior empirical research [106, 107]. From a computational perspective, image segmentation dates back to the seminal work of Minsky and Pappert [108] followed by several works in the 60s and 70s [109, 110]. At that time, segmentation was based on grouping elements as belonging to the same object. Adelson [111] proposed to group elements based on abstract textures and materials, advocating the idea of seeing *stuff* rather than *things*. This was the stepping stone for modern semantic segmentation, where the objective is to group the image regions based on labels with semantic meaning, without relying on individual objects [112]. Furthermore, the use of semantic labels transforms the clustering problem into a classification problem. Recent research using deep learning has gone one step further to produce *instance-aware* semantic segmentation [113]. In this case, we are back to the concept of seeing *things* by grouping pixels of single objects, but including a semantic label for the object.

For visually impaired people, basic scene understanding is essential for many everyday tasks and it also facilitates subsequent tasks of finer perception. In this work, we use segmentation to provide a basic visual representation of indoor scenes for prosthetic users. We combine both semantic segmentation and instance segmentation. We use instance segmentation to highlight relevant objects in the scene. This has a double purpose: on one hand, we are able to reduce visual clutter, which becomes indistinguishable noise in a low resolution implant array; on the other hand, the grouping highlights the silhouette of the object, making it more distinguishable. One of the main problems of using object silhouettes for recognition is the lack of sense of scale or perspective. Thus, we rely on a second semantic segmentation component to extract structural informative edges of the scenes, such as wall and ceiling intersections. Those edges provide an intuitive representation of the 3D structure of the room as concluded in [114], where it is shown that the results with the structural edges are significant and better than the results obtained without edges for scene recognition. The idea of combining instance and

semantic segmentation has been previously studied in the computer vision literature with different approaches [115, 116] and it has shown to be of great benefit for holistic scene understanding [117]. The limiting case where every pixel of the image has a semantic label and instance id is called panoptic segmentation [118].

Current state of the art methods for image segmentation are mostly based on deep neural networks [112]. Most recent developments of semantic and instance-aware segmentation are based on Fully Convolution Networks (FCN) [112, 113, 119–121]. A FCN is an architecture based on convolutional layers with added upsampling layers with skip connections to allow for detailed pixel prediction on arbitrary-sized inputs [122]. Similar approaches, like the U-net architecture, are able to provide accurate pixel prediction [123]. In this work, we use two different types of FCN-based segmentation to highlight the information available in the image and to present the most useful information to the user: PanoRoom [119] for semantic segmentation of structural elements and Mask-RCNN [120] for instance segmentation of relevant objects.

We evaluate and compare the proposed semantic and structural image segmentation with baseline methods through a Simulated Prosthetic Vision (SPV) experiment, which is a standard procedure for non-invasive evaluation using normal vision subjects [37, 38, 47–49, 54, 55, 64, 67, 69, 75, 95–101]. The experiments included two tasks: object recognition and room identification.

2.2 Methods

2.2.1 Subjects

Eighteen subjects with normal vision volunteered for the formal experiment. The subjects (four females and fourteen males) were between 20 and 57 years old.

Ethics Statements

The research process was conducted according to the ethical recommendations of the Declaration of Helsinki and was approved by the Aragon Autonomous Community Research Ethics Committee (CEICA) that evaluates human research projects, human biological samples or personal data. The research protocol used for this study is non-invasive, purely observational, with absolutely no-risk for any participant. There is no personal data

collection or treatment and all subjects were volunteers. Subjects gave their informed written consent after explanation of the purpose of the study and possible consequences. The consent allowed the abandonment of the study at any time. All data were analyzed anonymously.

2.2.2 Stimuli

We use a two step process to generate the stimuli used in the experiments. First, we process the original color image with the different methods stated in the following subsections. This generates a grayscale or binary image which corresponds to the signal to activate the electrodes of the retinal implant. Then, we use *simulated prosthetic vision* by generating the phosphene pattern as an image in a computer screen. The phosphene simulation has been designed to represent the descriptions of phosphene perception reported by retinal prosthesis patients.

The following sections describe the three segmentation methods used in the experiments to process the input images and generate the activation of the phosphenes. First, our proposal based on semantic segmentation with artificial neural networks (SIE-OMS). To our knowledge, this is the first work that have used deep learning models in this setup. Therefore, we have used two standard processing methods as baselines: a) detecting the silhouettes and structure within the scene with a standard edge detector (Edge), and b) generating the stimulus directly from the input image luminance (Direct). Examples of the resulting effect are shown in Figure 2.1. For reproducibility purposes, all the stimuli images used in the experiments can be found in the dataset available online¹.

SIE-OMS

We propose to combine two FCNs to select and highlight informative elements in indoor scenes as an intelligent way of activating the phosphenes. Specifically, we extract structural informative edges (SIE) and object masks and silhouettes (OMS) to later combined both, SIE and OMS, to build our proposed schematic representation of the scene (SIE-OMS), as can be seen in Figure 2.2. This idea comes from our previous study, where the results concluded that the representation of SIE in the schematic representation of the scene is significant and produce better results in object and scene recognition for SPV than the

¹Image dataset: <https://doi.org/10.6084/m9.figshare.11493249.v4>

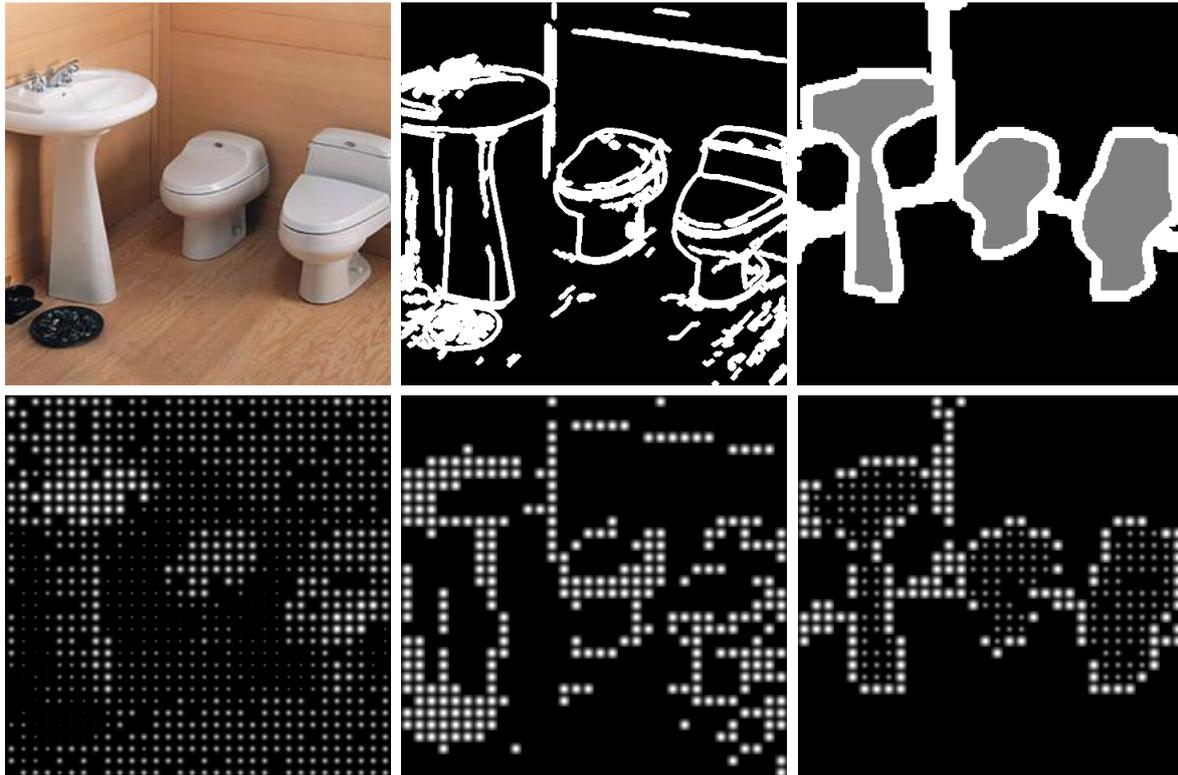


Figure 2.1: Top row: Example of a bathroom scene with the three processing methods used in this work (a) Direct image, (b) Edge image and (c) SIE-OMS image. Bottom row: the three processing methods in the SPV.

schematic representation without edges [114].

Structural Informative Edges (SIE) One of the main problems in the recognition of scene elements based on silhouettes is the lack of sense of scale or perspective. The scale and the structure of the scene can be achieved by detecting the structural informative edges (SIE), that is, those main edges formed by the intersection of the walls, floor and ceiling of the room. These edges can be seen in Figure 2.3. Our approach is based on the model by Fernandez-Labrador et al. [119] for indoor scenes. Similarly to the object masks network described below, this method is also based on a FCN for pixel classification. In this case, the network was trained to estimate probability maps representing the room structural edges, even in the presence of clutter and occlusions. The architecture of the network is an the encoder-decoder structure [119]. The encoder is built from a ResNet-50 model [124], pre-trained on the ImageNet dataset, with the final layer replaced with a decoder that jointly predicts layout edges and corners locations already refined. The output of the model is an unique branch whose output has two channels, corners and

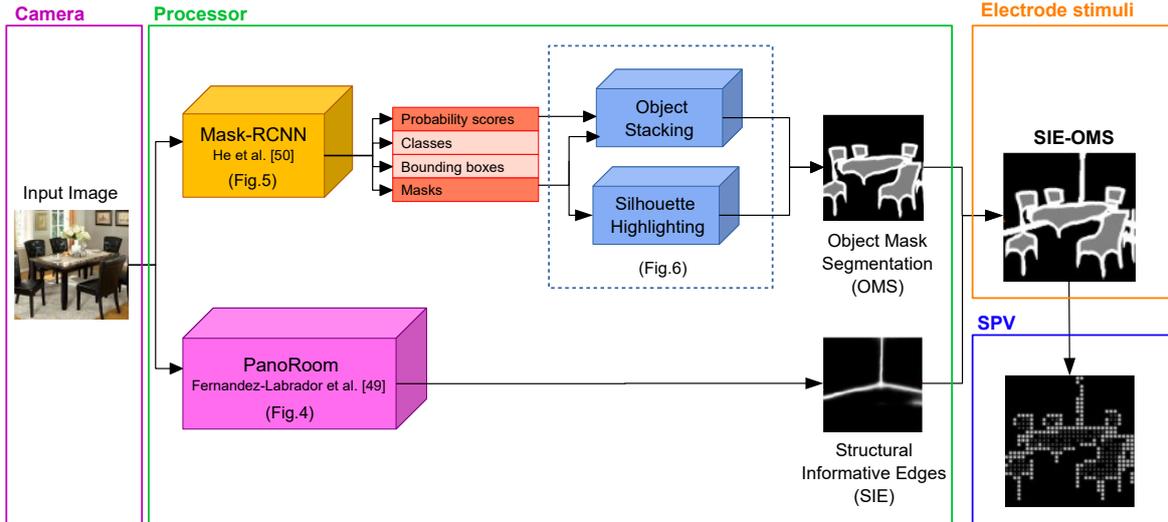


Figure 2.2: The stimulation of the electrode array is based on two information pathways to extract the regions of pixels that represents important objects (OMS) and structural edges (SIE). The regions are computed using two different types of FCN from He et al. ([120]) and Fernandez-Labrador et al. [119].

edges maps. In the decoder, the model employs skip-connections from the encoder to the decoder concatenating ‘up-convolved’ features with their corresponding features from the contracting part. In order to improve the training phase, Fernandez-Labrador et al. suggest to perform preliminary predictions in different resolutions which are concatenated and feed back to the network. The loss function for training is a pixel-wise sigmoid cross-entropy, regularized by the L1-norm of the network parameters. The loss function was minimized by using Adam with an initial learning rate of $2.5e^{-4}$ and exponentially decayed by a rate of 0.995 every epoch. They also applied 0.3 dropout² rate and $5e^{-6}$ weight decay and a batch size of 16 which allowed the learning of more complex rooms. In this work, we have used a model pre-trained with the LSUN dataset [125].

Object Mask and Silhouettes (OMS) We perform instance segmentation of objects using the original architecture of Mask R-CNN [120] which is partially represented in Figure 2.4. This method is an extension of Faster R-CNN [126] with several improvements and an extra branch to segment the object masks. The first part of the network, called a Region Proposal Network (RPN), proposes object bounding box candidates on the input image. These candidates are called regions of interest (ROIs). It also generates feature maps from the whole image. In our case, we used the Feature Pyramid Network (FPN)

²In this context, *dropout* refers to the technique used in deep learning to prevent overfitting. Do not confuse with the dropout of phosphenes.



Figure 2.3: Using [119] we detect the main structure of the room extracting the structural informative edges (SIE) (right) which are those formed by the intersection of walls, ceiling and floor of the room (middle).

[127]. The second module is a RoIAlign layer that pools a small feature map for the object region from those extracted by the FPN. Then, it aligns each ROI to the feature map. This is the *backbone* architecture. Then, the model splits in two branches, as can be seen in Figure 2.4. The *box branch* is based on the classification component of Faster R-CNN. It generates two outputs for each ROI: a) the class of the object present in the ROI and, b) a refined object bounding box using a regression model. The *mask branch* is a convolutional neural network that takes the high probability regions selected by the ROI classifier –box branch– and generates a binary masks of the object. Then, it uses upsampling and deconvolution layers to scale the predicted masks to the size of the ROI bounding box which gives the final masks, one per object. Regarding the loss function for the model, it is composed by the total loss in doing classification, generating bounding box and generating the mask. The mask loss was defined only on positive RoIs and the mask target was the intersection between an RoI and its associated ground-truth mask. The training was performed with a batch size of 16 and for 160k iterations. The learning rate was 0.02 which was decreased by 10 at the 120k iteration. They also used a weight decay of 0.0001. For this work, we have used a pre-trained model on the COCO dataset [120, 128]. Thus, we have only considered the object classes that were already defined in the pre-trained model. In order to speed up computation and remove spurious detections we removed the object classes of clearly small objects (e.g.: scissors, banana, etc.) as the scale of the scene in the image would not have allow it to be identified, and non-indoor objects (e.g.: car, tree, etc.). Once the object masks have been generated by the network,

we highlighted the contour of each mask to avoid confusion on overlapping masks, as can be seen in Figure 2.5. We also performed morphological operations to reduce the aliasing effect when translated to phosphene images.

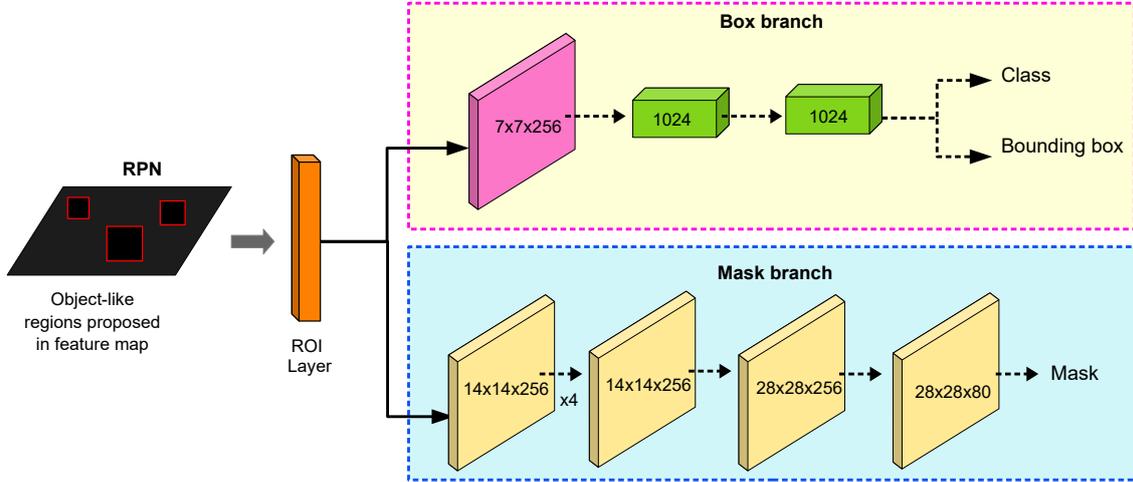


Figure 2.4: Above: box branch for classification and bounding box regression. Below: mask branch for predicting segmentation masks on each Region of Interest (ROI). Numbers denote spatial resolution and channels. Arrows denote either convolutions, deconvolutions, or fully connected layers. The $x4$ means 4 consecutive convolution layers. (Adapted from He et al. [120])

Dealing with occlusions Although this algorithm has achieved good results for object segmentation, there are more complicated cases, such as images with overlapping objects or scenes with occlusions, where the view of one object may be blocked by other objects. In that case, we could use a depth sensor, such as an RGB-D camera, or a stereo camera to estimate the depth. Alternatively, there are some works to estimate the depth purely, based on monocular information [129]. As a proof of concept, we found that the probability score for the detection network was correlated to the level of occlusion of each object. In non-occluded objects its form is complete and therefore its recognition is more likely detected. In contrast, the form of the occluded objects is not complete and therefore its recognition is less likely to be detected. That is, a high probability is most likely to appear in objects that are in the front. Thus, we stacked the instances from the least to the highest probability, leaving the objects with the highest score overlapping the objects with the least score, as can be seen in Figure 2.5. This was confirmed experimentally for our setup, which is a simple problem with a limited number of classes and scenarios. In the case of a more general environment with more classes it would be necessary to use a

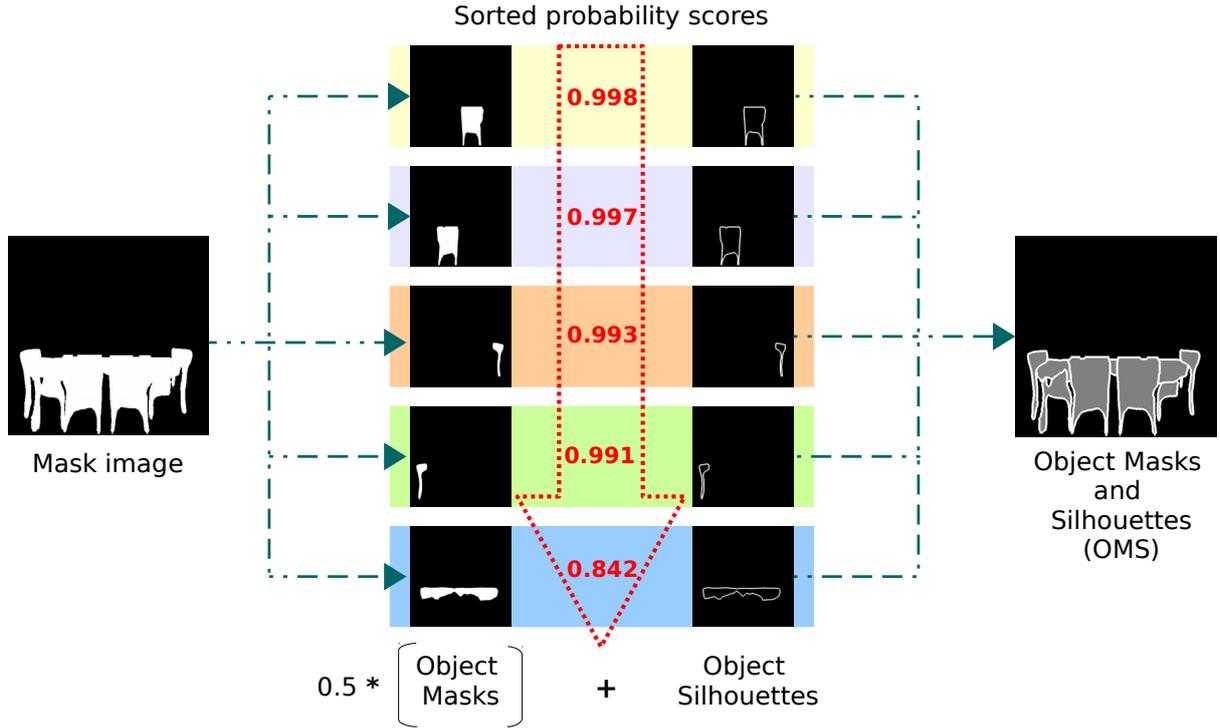


Figure 2.5: Object masks were generated from [120] and were sorted by probability scores to avoid occlusions between objects. The extracted information was combined in an image highlighting the silhouettes of the objects in white with the object masks in gray.

more complex model, but that is beyond the scope of the paper.

The final representation of the SIE-OMS method is a superposition of both parts, SIE and OMS, always assuming the edges as background and object masks as foreground.

Baseline methods

We have considered two baseline methods that are the most used in the literature and that follow a completely different structure to our SIE-OMS model [62, 130–132]. We compared SIE-OMS with two baseline methods used in retinal prosthesis: a) a direct method that converts the input image directly to the phosphene map by averaging the brightness on the region covered by each phosphene, and b) a standard edge detector to extract brightness contours (see Figure 2.1). The direct method has proved to be very effective in scenes where high contrast predominates [130]. Edge detectors have also been previously used for prosthesis vision and phosphene images [62, 131, 132]. Since the contours of an image holds much information, edge extraction is a useful method of encoding and selecting the information contained in an image. The drawback here is that the understanding of a

complete scene in low vision represented by edges may be more challenging because the amount of clutter. For example, Sanocki et al. investigated how complicated is an edge extractor method comparing object recognition with and without removal of background clutter with edge images [133]. The results showed that the increase in the number of edges greatly increase the complexity. For the edge detector, we used the Canny implementation from the `scikit image` Python package with the default parameters [134]. In this case, we also added morphological operations (dilation) to reduce aliasing without adding clutter.

Phosphene simulation

As commented before, first, the input images are processed by one of the three methods (SIE-OMS, Direct, Edge) resulting in the grayscale images from Figure 2.1. Then, these grayscale images are used to activate the phosphene map. In this work, we have used a simulated phosphene map on a computer screen, but the same activation images could be directly applied to the retinal implant.

Based on previous studies with simulated prosthetic vision [45, 135, 136], we approximate the phosphenes as grayscale circular dots with a Gaussian luminance profile –each phosphene has maximum intensity at the center and gradually decays to the periphery, following a Gaussian function–. The intensity of a phosphene is directly extracted from the intensity of the same region in the processed image. The size and brightness are directly proportional to the quantified sampled pixel intensities. For our phosphenic images, the array of phosphenes was limited to 32 x 32 (1024 electrodes) and 8 different luminance levels according to the number of luminance levels attainable in human trials using retinal prostheses [45, 47]. We also included a 10% dropout of electrodes, which is a standard value used in the literature [136]. The dropout percentage has shown that significantly affects the performance of recognition tasks decreasing recognition accuracy as the dropout percentage increases [137]. The complete process of phosphene generation can be found in the Supplementary material (see Appendix).

2.2.3 Experimental setup

Most of the SPV configurations are usually based on a computer screen for the presentation of static or dynamic phosphene images [37, 138, 139]. This methodology allow controlled

evaluation of normally sighted subject response and task performance which is fundamental to know the way humans perceive and interpret phosphorized renderings. SPV also offers the advantage of adapting implant designs to improve the perceptual quality using image processing techniques without involving implanted subjects. In our case, the participants were normal sighted subjects seated on a chair facing a computer screen at 1m distance resulting in a 20 degrees simulated field of view, as can be seen in Figure 2.6.

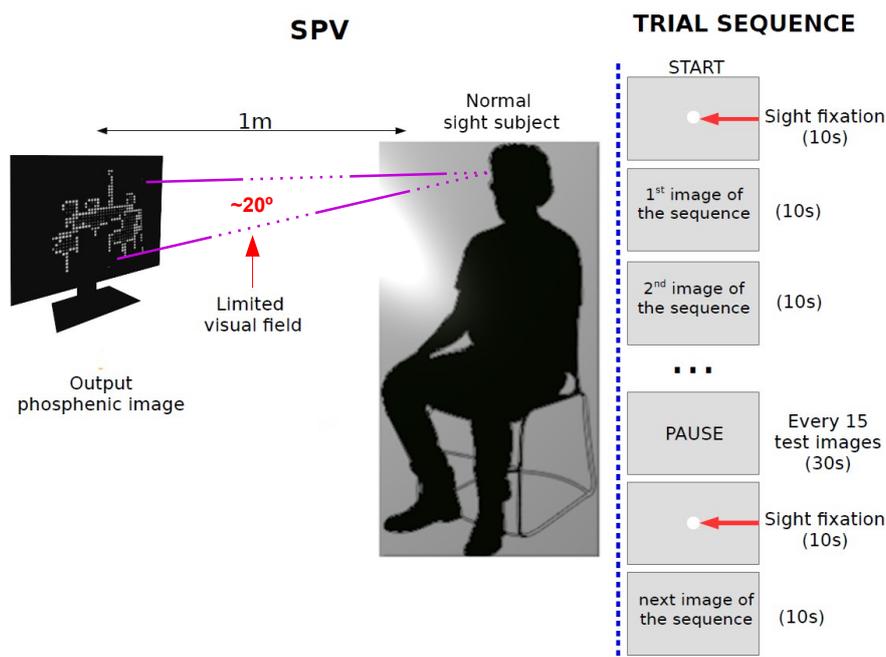


Figure 2.6: SPV setup: Subjects were seated on a chair facing a computer screen at 1m distance. The visual field was 20 degrees that simulates the prostheses device. Trial setup: Each gray rectangle represents the image shown on the computer monitor during the trial. Each image appeared for 10 seconds and switched for the next image automatically. Break time between image sequences was 30 seconds. The complete experiment took approximately 15 minutes.

For the formal experiment, subjects were recruited to complete two tasks: object recognition and room identification. The recognition accuracy was analyzed after the trials. Each trial consisted of a sequence of images presented randomly to the subject with the proposed SIE-OMS stimuli method and the two baseline methods (Edge, Direct), as can be seen in Figure 2.7. At the beginning of the experiment, a white dot was displayed in the center of the screen indicating where the subjects had to maintain the fixation sight until the beginning of the task. Next, each phosphoric image appeared for 10 seconds and switched for the next image automatically. This procedure was repeated for the

other test images. To avoid distractions in the participants, they verbally indicate the type of objects seen in each image and their selection of room type keeping the fixation sight on the screen. The responses of each image were annotated by the experimenter. If the subjects did not respond within the 10 seconds that the image is displayed, the result of the test image was considered not answered (NA). If the subjects were only able to respond to one of the two phases of the experiment, only the unanswered phase was considered as not answered. Every 15 test images we made a pause of 30 seconds. The complete experiment took approximately 15 minutes.

The experiments were conducted using a public database of indoor scenes [140]. All the images from the database are still life scenes, from arbitrary scenarios, locations, clutter, cameras and lightning conditions. Some images are from old phone cameras with very poor quality and resolution to be more challenging as a computer vision benchmark. Thus, we replaced some images with the first results of querying Google Images with the room label, that also matched the database features (e.g., still life, mid-wide view...). For each of the six categories, we randomly selected 50 images. Hence, we conducted the experiment using 300 images from different indoor environments. The original images were processed using our proposed method and the two baseline approaches, resulting in 900 phosphoric images. Prior to beginning the experiment, subjects were informed about the number of images in the experiment (54 images per subject). Subjects were unaware that multiple image processing strategies were used in the experiment, although a screen with four images were shown to the subjects at the beginning of the test. These demo images were not included in the experiment, to avoid learning effects. Subjects were informed that all scenes were indoor scenarios, but they were not informed about the types of room, neither the object classes, nor the number of objects in each image. The types of room studied were: *bathroom, dining room, living room, kitchen, office and bedroom*. No subject identified a type of room or scene not belonging to that list. In most of the tests, the objects identified by the subjects were: *chair, table, couch, toilet, bath, sink, bed, oven/microwave, refrigerator, laptop*. This coincides with the list of classes used for our SIE-OMS which was selected without looking at the database and before conducting any trial or test. As commented before, the object classes were those already included in the pre-trained model. However, in two images with the direct method, a couple of subjects were able to find a *window* that our system did not detect because the

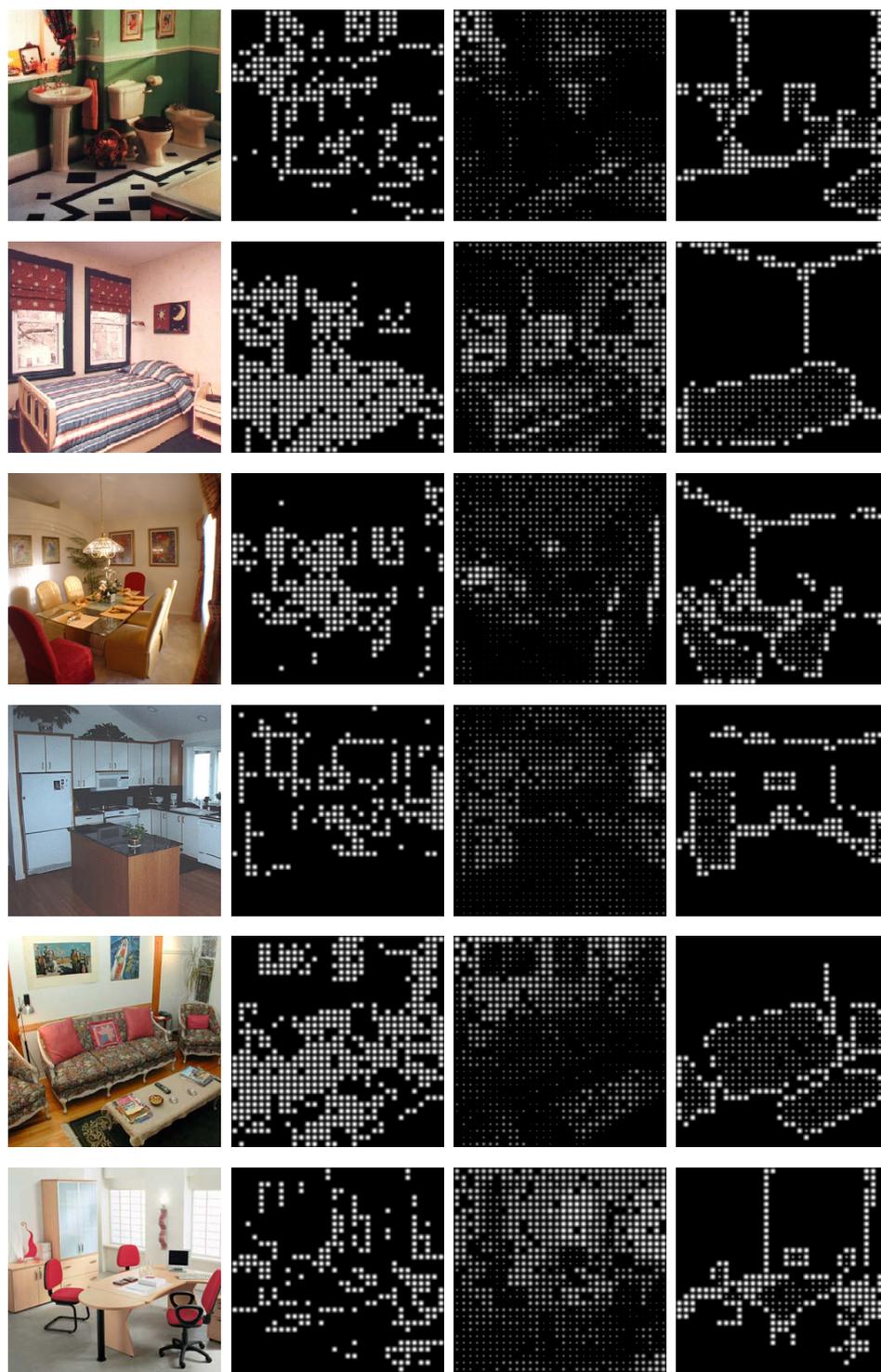


Figure 2.7: Six examples of indoor environments represented with 1024 phosphenes (rows: bathroom, bedroom, dining room, kitchen, living room and office, respectively). Each column shows: a) input images, b) images processed using the Edge method, c) images processed using the Direct method and d) images processed by our SIE-OMS method, respectively.

class was not included. Furthermore, in a couple of cases a subject wrongly identified *wardrobe* and *door* in images containing a *fridge*.

2.3 Results

The following section shows the results of the experiment. We analyze separately the results of object recognition phase and room identification phase. We show the percentage of correct responses in both tasks and we include 95% confidence intervals. We also differentiate between incorrect response and no answer.

2.3.1 Comparison of stimuli generation methods

Table 2.1 and Figure 2.8 show the global results for object recognition and room identification tasks considering the proposed stimuli generator (SIE-OMS) and the two baseline methods (Edge and Direct). The analysis of the average correct responses for both tasks reveals a significant difference between methods ($p < 0.001$). In both tasks, the results show a considerably better performance of SIE-OMS compared to the other methods. The SIE-OMS method has the highest percentage of correctly identified objects (62.78%) compared to Edge (19.17%) and Direct (36.83%) methods. Likewise, there is a clear increase in the percentage of success in the room identification of SIE-OMS versus Edge and Direct method. The number of unanswered responses for our method was also smaller, indicating that there was no difficulty in the comprehension of most of the images. In contrast, it is worth noting the high percentage of unanswered responses for the Edge method, reaching more than 70% of the scenes.

Table 2.1: Global object recognition (OR) and room identification (RI) values for each phosphenic stimuli method. Comparison of mean responses and standard deviation grouped by type of phosphenic image method (Edge, Direct and SIE-OMS). 95% of confidence interval for the mean difference.

Method	%	% RI
Edge	19.17 ± 4.45	13.33 ± 3.85
Direct	36.83 ± 5.46	35.33 ± 5.41
SIE-OMS	62.78 ± 5.50	70.33 ± 5.17

Figure 2.9 and Figure 2.10 show the results for the object recognition and room

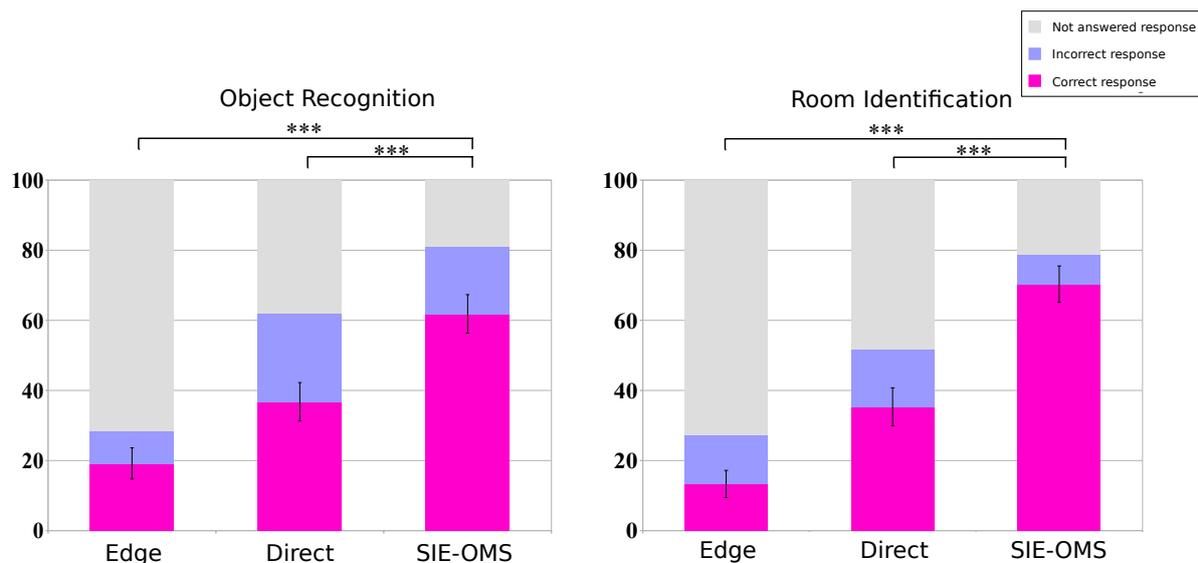


Figure 2.8: Percentage of correct, incorrect and not answered responses in a single trial. Higher scores in correct responses indicate that subjects were able to identify and recognize the objects and the type of room in each test image. Higher ratios of not answered indicate that subjects were not able to identify and recognize the objects and the type of room in each test image. The general findings are that: SIE-OMS method improves the identification of the objects resulting to be the most effective method. This translates in an increase in the number of correct answers for the room type identification test for the SIE-OMS method. Results also show that the Edge method is the least effective with the highest percentage of non responses images for the two tasks. The test found significant difference between SIE-OM and Direct method ($p < .001$). The same conclusion was found between SIE-OM and Edge method ($p < .001$). Where: ***= $p < .001$; **= $p < .01$; *= $p < .05$; ns= $p > .05$. All t-tests paired samples, two-tailed.

identification tasks for each room-type, respectively. As before, when comparing the baseline methods versus our approach, the highest number of correct responses is obtained for SIE-OMS method for all room types. Besides, the largest difference in results was obtained comparing the Edge method versus the SIE-OMS method ($p < 0.001$) for all room-types. However, there was no significant difference for *kitchen* type in Direct vs SIE-OMS ($p = 0.464$). Similarly, there is a significant difference for *living room* ($p < 0.05$), *office* ($p < 0.01$) and *bathroom* ($p < 0.01$). On the other hand, the results of room identification task for each room-type (Figure 2.10) provide additional support for the SIE-OMS method since this method also has the best percentage of correct responses in each room-type, exceeding 85% for the cases of bedroom and dining room. In the same way as in the identification of objects, the case of the *kitchen* obtained the worst results, followed by the *office* case. Taken together, these findings indicate that SIE-OMS method was significantly effective improving object recognition and room identification, yet also significantly more

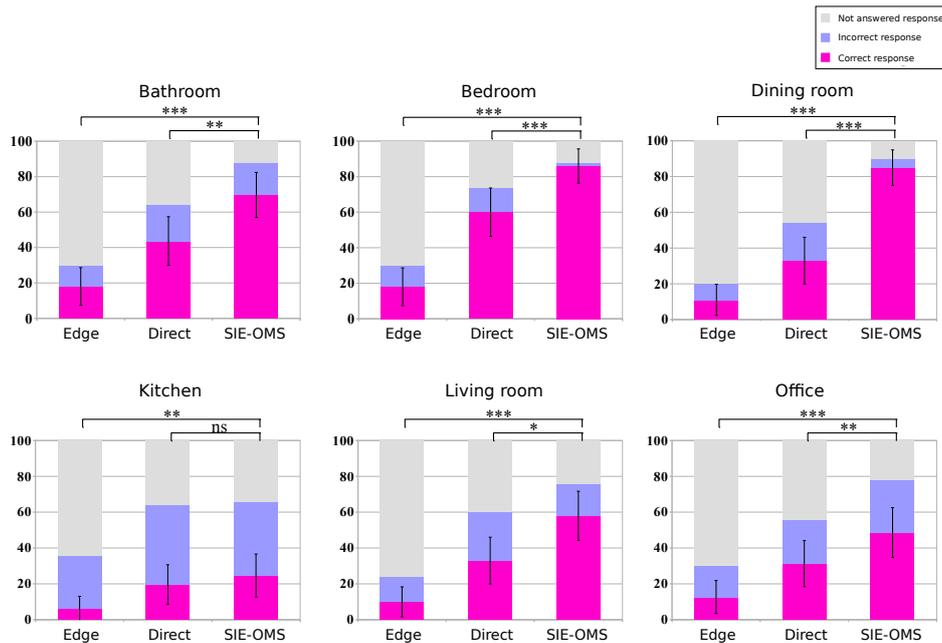


Figure 2.9: Higher scores in correct responses indicate that subjects were able to recognize the objects in each room. Higher ratios in non responses indicate that subjects were not able to recognize the objects in each room. The SIE-OMS method obtained the highest score of the three methods in all room types compared with Edge and Direct methods. The results also show how the most difficult room was the kitchen. ***= $p < .001$; **= $p < .01$; *= $p < .05$; ns= $p > .05$. All t-tests paired samples, two-tailed.

effective than the baseline methods, Edge and Direct.

Figure 2.11 shows four examples of failed and successful tests from the three methods. The two top rows show a *bathroom* and a *bedroom* scene where the identification of the objects and room was a success for all the methods. This is due to the location of a characteristic object with a clear silhouette in the center of the image that also helps in the identification of the room. Contrary, the bottom rows show a *kitchen* and an *office* where the recognition of the objects and the identification of the type of room failed in all cases as a result of the lack of distinguishable shapes (rectangle silhouettes) and visual clutter.

2.3.2 Performance analysis of SIE-OMS

We also analyzed the performance of the proposed SIE-OMS method. The SIE-OMS system detected all the clearly visible objects of the scenes and even most of the occluded objects that matched the selected classes. Structural edges also improve the performance of our method. Recovering the main structure of the room provide sense of scale or

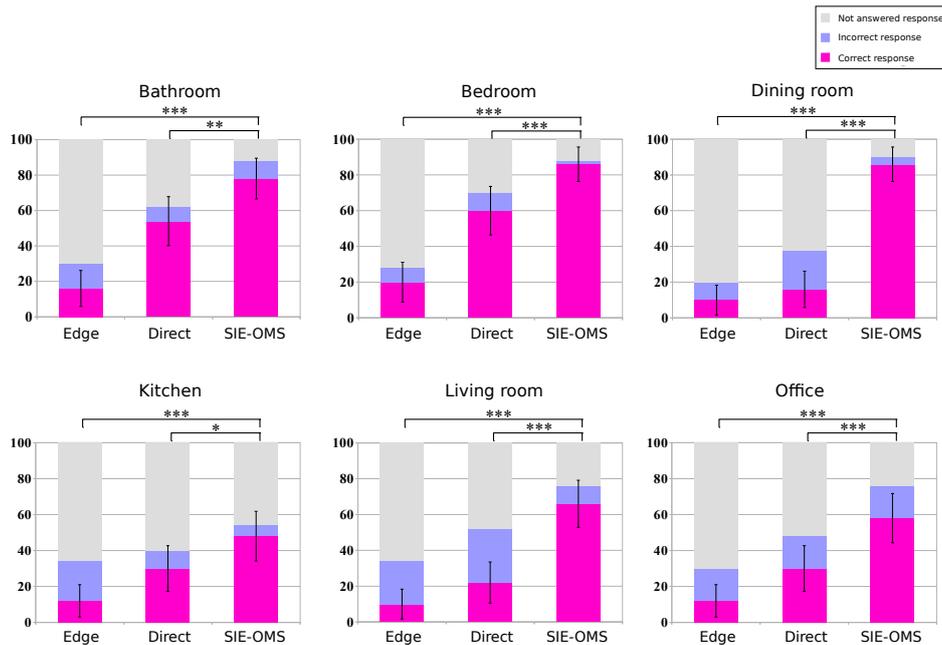


Figure 2.10: Higher scores in correct responses indicate that subjects were able to recognize the type of room in each test image. Higher ratios in non responses indicate that subjects were not able to recognize the type of room in each image. The SIE-OMS method obtained the highest score of the three methods in all room-type compared with Edge and Direct methods. In the same way as in the identification of objects, results also showed how the most difficult room was the kitchen. $***=p<.001$; $**=p<.01$; $*=p<.05$; $ns=p>.05$. All t-tests paired samples, two-tailed.

perspective of the objects and hence a better understanding of the 3D scene. Table 2.2 shows the confusion matrix of room-type based on answered images (correct and incorrect responses). Table 2.3 shows the confusion matrix of the room-type based on the total images of the test (correct, incorrect and no answer).

Table 2.2: Confusion matrix results for room identification based only on answered images (correct and incorrect responses) using SIE-OMS method.

Actual/Predicted	Bathroom	Bedroom	Dining room	Kitchen	Living room	Office	Total	Recall
Bathroom	0.89	0.00	0.00	0.00	0.09	0.02	1.00	88.64
Bedroom	0.00	0.98	0.00	0.00	0.02	0.00	1.00	97.73
Dining room	0.00	0.00	0.96	0.02	0.00	0.02	1.00	95.56
Kitchen	0.04	0.00	0.00	0.89	0.04	0.04	1.00	88.89
Living room	0.05	0.03	0.00	0.00	0.87	0.05	1.00	86.84
Office	0.13	0.08	0.00	0.03	0.00	0.76	1.00	76.32
Total	1.11	1.08	0.96	0.94	1.02	0.90	6.00	
Precision	80.02	90.28	100.00	94.82	85.21	85.01		

Concerning the performance of our method, the recall and the precision are high, reaching in some cases up to 97% (Table 2.2). The diagonal elements show the number of

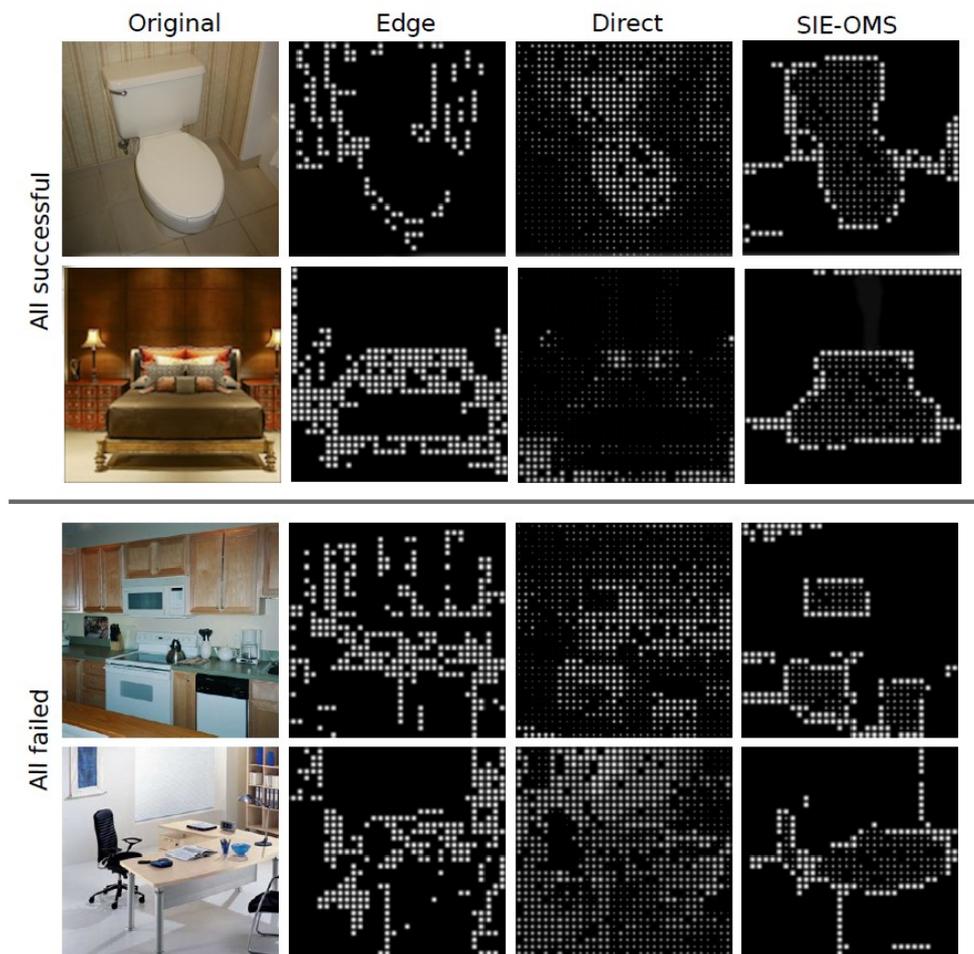


Figure 2.11: Some examples of phosphenic images generated with the three methods. Successful images (top rows) and cases of images failed by the subjects (bottom rows) with the three approaches: Edge, Direct and SIE-OMS, respectively.

Table 2.3: Confusion matrix results for room identification based on the total images (correct, incorrect and no answer (NA)) using SIE-OMS method.

Actual/Predicted	Bathroom	Bedroom	Dining.r	Kitchen	Living.r	Office	NA	Total	Recall
Bathroom	0.78	0.00	0.00	0.00	0.08	0.02	0.12	1.00	78.00
Bedroom	0.00	0.86	0.00	0.00	0.02	0.00	0.12	1.00	86.00
Dining room	0.00	0.00	0.86	0.02	0.00	0.02	0.10	1.00	86.00
Kitchen	0.02	0.00	0.00	0.48	0.02	0.02	0.46	1.00	48.00
Living room	0.04	0.02	0.00	0.00	0.66	0.04	0.24	1.00	66.00
Office	0.10	0.06	0.00	0.02	0.00	0.58	0.24	1.00	58.00
Total	0.94	0.94	0.86	0.52	0.78	0.68	1.28	6.00	
Precision	82.98	91.49	100.00	92.31	84.62	85.29			

correct classifications for each class. Hence, most of confusions are found in bathroom, living room and office. Office was confused with bathroom because of the similarity of

shape between some chairs and toilets. In addition, office was confused by bedrooms since many of them usually have study desks in the bedrooms. There were other less relevant cases where the dining room was confused with an office since both are composed of chairs and tables. This confusion can be explained because the database is from Northamerican locations, while the subjects live in Spain where apartments commonly join the dining room and the kitchen.

Note that when the unanswered responses are taken into account (Table 2.3), the recall for the kitchen case decreased significantly (from 88.89% to 48.00%). This means that the kitchen room is more difficult to be identified. This low performance in the kitchen identification is mainly because the information provided turned out to be very limited in this case. For instance, ovens, microwaves and fridges with a rectangular shape masks were sometimes confused with windows, doors or wardrobes (which are object classes not considered by our system).

2.4 Discussion

The visual information in interpretation of the phosphene simulation is an important issue due to the limited capabilities of retinal implants. Low resolution, limited dynamic range and narrow visual field are some of the limitations present in current retinal prostheses [45, 85]. Furthermore, Nanduri et al. [84] showed that phosphenes are not perfectly located in the visual field corresponding to a specific grayscale pixel. Electrically elicited phosphenes change in form and size with increasing amplitude. However, depending on the type of device the perceptual distortions will be affected differently. For example, in retinal prostheses this distortion produce visual effects called *comets* that might result in a substantial loss of information [141]. Other devices, such as optogenetic technologies, may suffer a loss of temporal resolution, while cortex implants suffer from crosstalk [141]. This fact results in loss of visual information which affects patient perception. However, there are research groups using computer vision approaches to try to expand the perceptible visual field in implanted patients to provide useful information in the peripheral vision [142, 143].

Another important limitation of retinal implants is phosphene dropout, which has been reported in retinal prostheses trials [93, 144] as a result of very high threshold values needed to elicit phosphenes in areas with a high number of degenerating nerve cells.

Clinical trials by Thompson et al. [63] indicated that the dropout rate has significant effects on the speed and accuracy of recognition tasks. Similarly, Cao et al. [145] showed that the accuracy and efficiency in writing tasks decrease as the variability of distortion and dropout percentage increase.

To overcome the limitations of implants, SPV researchers have tried to optimize the image presentation to deliver the effective visual information in daily activities. For example, Vergnieux et al. [99] limited the cues in a virtual scene using different renderings methods, highlighting structural cues such as the edges of different surfaces for navigation. For the same purpose, Perez et al. [101] proposed a phosphene map coding using a ground representation of the obstacle-free space and a ceiling representation based on vanishing lines. Wang et al. [47] proposed two image representation strategies using background subtraction to segment moving elements for object recognition. Similarly, Guo et al. [97] and Li et al. [67] proposed two image processing strategies based on a saliency segmentation technique. For scene recognition, McCarthy et al. [146] presented a visual representation based on intensity augments in order to emphasise regions of structural change.

In terms of complex scene understanding, just few SPV studies have been proposed [67, 147]. It is well established that in realistic environment, which is made of complex scenes, the observer is forced to select relevant elements [148]. That is necessary to quickly understand the meaning of a scene as well as for object search. For instance, the set of objects in the environment give rise to a corresponding set of representations in the observer. Each representation describe the identity, location, and meaning of the item it refers to finally forming a literal representation of the environment. Some research on the visual perception of subjects has shown that because the fixation of the gaze changes in a short period of time when an environment is observed, the content of the scene can not be integrated into a complete and detailed representation [149, 150]. This suggests that such complete and detailed representations are not needed to obtain a meaning of the scene. Just a few set of object and scene elements are enough to provide access to semantic information [151].

A well-known result in psychophysics highlight that grouping elements in a scene are fundamental for scene understanding [107]. First, the grouping of pixels in a region defines a contour. In many cases, shape alone permits recognition of objects. Biederman

et al. [149, 152] demonstrated that the silhouettes of the objects are generally very easy to identify and to recognize. The silhouette conveys only part of the visual information needed for the interpretation of an object. Concretely, the concepts such as convexities, concavities, or inflections of contours allow the observer to infer the surface geometry [153]. However, this bottom-up perception can be computed first and to help any top-down search to converge to the right answer. This can help to understand the visual scene through the interpretation of its content. However, in order to fully understand a scene, it is not only important the identification of individual objects comprising the scene but also their relative locations and relations [151]. Based on this idea, the segmentation of the scene into elements with semantic meaning becomes a key point in low vision.

The state of the art in semantic segmentation include deep learning algorithms. Specifically, FCNs have proven to be successful in various recognition tasks such as semantic segmentation of images. In this work, we use two FCNs to select and highlight useful information in indoor scenes such as relevant object masks and silhouettes and structural edges which recover the main structure of the scene providing sense of scale or perspective of the objects. Even though deep learning methods are known for being resource-hungry during training, they can achieve real time performance for prediction even in mobile or embedded devices [154]. Thus, our method could be easily integrated in an implant device.

The performance of the proposed visual stimuli, the SIE-OMS method, was investigated for object recognition and room identification tasks. We introduced the effect of dropout with a 10% of phosphenes omitted at random. This effect has been shown in other studies that decreases the performance of subjects in daily tasks, although it is known that with practice it will improve performance [63, 84, 141, 143]. Our results show that generating phosphene images by extracting specific segments of the scene such as structural informative edges and objects shapes are effective at improving object recognition and room identification. Moreover, the SIE-OMS method produces a large improvement on object recognition and room type identification compared to standard methods in SPV. Here, we have taken the pre-trained neural network model of He et al. [120] with the same classes as it had pre-defined without modifying any of them. The pre-defined classes coincided with those classes detected by users with the Direct method, which does not depend on the model of [120], since both methods, the SIE-OMS and the Direct, are

independent. Note the case of the “window” class that was detected by users with the Direct method but was not a pre-defined class in the model [120]. We consider large objects since the scale of the images allowed a complete view of the room and that could also be identified by the users with the Direct method. However, object appearance alone is not enough for accurate object recognition in certain scenes. Since the only piece of visual data that our system uses for each object is its shape, the introduction of complementary information such as the object label could make recognition easier and avoid confusion between objects with similar shape. These factor will be considered in future studies for more realistic practices. We also note that structural edge detection is fundamental for performing tasks such as self-orientation and building a mental map of the environment. Finding such structure is crucial for personal mobility with retinal prosthesis, where the bandwidth of image information that can be represented per frame is quite restricted. Overall, we can affirm that the perception and comprehension of the scene can be obtained with just a few set of elements represented in the environment.

2.5 Conclusions

We have propose a new visual representation of indoor environments for prosthetic vision, which emphasizes the scene structure and object shapes. By combining the output of two FCN for structural informative edges and object masks and silhouettes, we have demonstrated how different scenes and objects can be quickly recognized even under the restricted conditions of prosthetic vision. Our results demonstrate that our method is well suited for indoor scene understanding over traditional image processing methods used in visual prostheses. The key idea of our current results is that, with only a few significant elements of the scene, it is possible to obtain a good perception of the environment, even in complex and occluded scenes. We believe that in the near future, deep learning-based algorithms may aimed at improving a patient’s scene understanding.

2.6 Related Publications

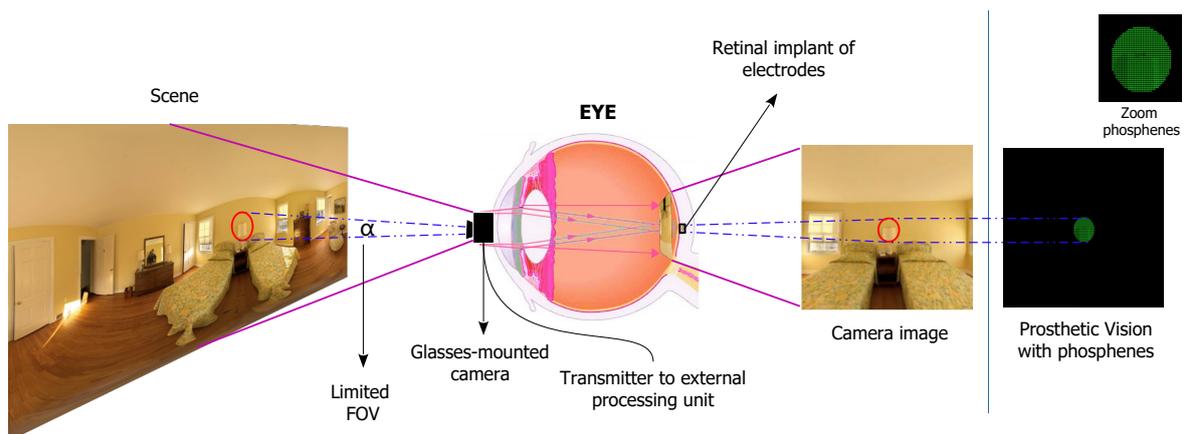
1. Sanchez-Garcia M, Martinez-Cantin R and Guerrero J. J, 2020. “Semantic and structural image segmentation for prosthetic vision”, *PLoS ONE*
2. Sanchez-Garcia M., Martinez-Cantin R., and Guerrero, J. J, 2019. “Indoor scenes

understanding for visual prosthesis with fully convolutional networks". In *14th International Conference on Computer Vision Theory and Applications*

Chapter 3

3 Influence of field of view

Visual prostheses are designed to restore partial functional vision in patients with total vision loss. Retinal visual prostheses provide limited capabilities as a result of low resolution, limited field of view and poor dynamic range. Understanding the influence of these parameters in the perception results can guide prostheses research and design. In this work, we evaluate the influence of field of view with respect to spatial resolution in visual prostheses, measuring the accuracy and response time in a search and recognition task. Twenty-four normal participants were asked to find and recognize usual objects, such as furniture and home appliance in indoor room scenes. For the experiment, we use a new simulated prosthetic vision system that allows simple and effective experimentation. Our system uses a virtual-reality environment based on panoramic scenes. The simulator employs a head-mounted display which allows users to feel immersed in the scene by perceiving the entire scene all around. Our experiments use public image datasets and a commercial head-mounted display.



3.1 Introduction

Retinitis pigmentosa and aged-related macular degeneration are the two major retinal degenerative diseases that cause a loss of vision [79, 80]. These diseases involve gradual loss of photoreceptor, or rod cells, while generally preserving the inner retinal cells. This results in a progressive loss of vision. Retinal prostheses are a promising technology to improve vision in patients with such advanced degenerative diseases [25, 81–83]. These visual prostheses can partially restore vision, bypassing damaged photoreceptors and electrically stimulating the surviving retinal cells, such as the retinal ganglion cells [155]. The Argus II epiretinal prostheses (Second Sight Medical Products Inc., Sylmar, CA, USA) is the most widely used retinal prostheses world-wide [25, 81]. It is made up of an external component (glasses-mounted camera) and an implanted component (electrode implant). The camera acquires images from the real-world that are transmitted to a portable visual processing unit linked to the camera. The processed information is sent to the retina via electrical impulses in the implant by an electrode array. The stimulation can activate a group of neurons in a small localized area of the retina leading to percepts of spots of light known as “phosphenes” [156]. The brain interprets patterns of phosphenes in the restricted area as visual information. Results in implanted subjects have demonstrated partial visual restoration, with improvement in both coarse objective function and performance of everyday tasks [157–166].

There are still physiological and technological limitations of the information received by implanted patients. The number of electrodes and implant size limit the maximum amount of information that can be provided by the stimulating array. This fact has restricted the degree of visual resolution (up to 1500 phosphenes) and dynamic range of the visual perception (8 gray levels) that can be delivered to the user. Depth perception is also not possible due to low resolution or monocular implantation. Although most current implants have shown good results using static electrode stimulation, they only elicit perception of multiple isolated phosphenes, sometimes resulting in a combination of non-coherent shapes. Therefore, alternative approaches using dynamic activation of an electrode sequence are being studied [167]. Additionally, field of view (FOV) is a key limitation affecting visual experience of recipients. The ocular anatomy and surgery are two major limiting factors for the narrow FOV [142, 168]. Current systems provide a FOV

of approximately $18^\circ \times 11^\circ$ in the retinal area, which correspond to the FOV subtended by the electrode implant on the retina.

The constrained FOV limits mobility and recognition capabilities of the prosthetic vision system reducing quality of life [169–171]. To improve the design of current retinal prostheses in terms of FOV and resolution, one of the questions to be answered is whether expanding the FOV of the input image can benefit implanted subjects. Some studies have attempted to improve the narrow FOV on the clever design of microelectrode arrays [172, 173]. Lohmann et al. [172] developed a flexible and thin retinal implant consisting of a multielectrode array approximately three times the size of the comparable epiretinal Argus II device which shows the possible recovery of meaningful peripheral vision. Other devices for retinitis pigmentosa patients (tunnel vision) have been designed to expand the projected FOV on their retina, for example, minimizing the scene zooming out using a camera [174, 175]. He et al. [176] studied the influence of FOV in the Argus II system with thermal imaging by changing the mapping (zoom out) between the sensor and the electrode array. Zooming out turns out in a larger visual angle mapped onto the same implant region on the retina, effectively decreasing the spatial resolution of their prosthetic vision. Alternatively, Ameri et al. [173] have designed an implant with wider FOV, by spreading the electrodes in a wider retinal surface. This setup seem to have advantages for motion perception and head scanning, at the expense of a reduced electrode density and fine detail. Therefore, the optimal implant area for a fixed number of electrodes remains an open question, which we address in this study.

One of the concerns that has limited the development and wider use of retinal prosthetic devices is how to evaluate their utility and function in terms of benefit to the patient and, consequently, how to predict in which direction to develop these devices [32]. Most studies in this field have incorporated performance-based measures and questionnaires to try to understand the relative importance of visual parameters such as resolution, FOV and visual acuity, in the performance of daily tasks in subjects with visual impairment. However, studies with implanted subjects have limited statistical power and require cumbersome experiments. For example, the study of He et al. [176] was limited to four implanted participants. Simulated prosthetic vision (SPV) opens the opportunity to evaluate potential and forthcoming functionality, in early stages of design, of these implants with larger studies by using normally sighted participants. SPV is a standard procedure

for non-invasive evaluation using participants without visual impairments. Furthermore, SPV systems allow for quick or even interactive modifications of the parameters of the simulation.

Researchers have previously used SPV for analysis of the visual perception in terms of resolution or FOV with normally sighted subjects. For example, Fornos et al. [177] used SPV to study how such restrictions of the amount of visual information provided would affect performance on simple pointing and manipulation tasks. Li et al. [67] applied an image processing algorithm to the image-to-electrode mapping process which improved the ability of the prosthesis visual perception under SPV. Hayes et al. [95] designed a set of tasks to assess performance of object recognition and manipulation and reading using different sizes of electrode array. Contrary to our approach in this paper, they used a constant electrode density, meaning that wider arrays also had larger number of electrodes. In a subsequent study, Dagnelie et al. [38] explored minimal visual resolution requirements of a simulated retinal electrode array for mobility in real and virtual environments, experienced by normally sighted subjects in video headsets.

These experiments relied on some form of head-mounted device which allowed a virtual-reality experience during the experiment, but the *immersiveness* of experience was poor due to the technological limitations compared to modern commercial VR systems. Furthermore, most SPV studies use computer screens to present phosphoric images [37, 77, 138]. Modern commercial VR systems have specifications in terms of resolution, latency and response time that allows a fully immersive experience with unexpensive equipment and computers. Using these devices, subjects are immersed and able to interact with complex environments. SPV with commercial VR systems, such as the Oculus Rift used in this work, can be a useful tool to evaluate everyday tasks in more realistic setups and complex simulations. Furthermore, prosthetic vision can be assessed in controlled, real or simulated environments.

In this work, we evaluated the influence of FOV and resolution of the prostheses on the subject's performance in a recognition task, since it is of high priority for patients with visual diseases such as retinitis pigmentosa. Concretely, we analyzed an object search and recognition task performance in indoor scenes with different reduced FOVs and resolutions limited to hundreds of electrodes. For that, we present a new VR system for more realistic SPV environments using panoramic scenes. The VR system could be extended to more

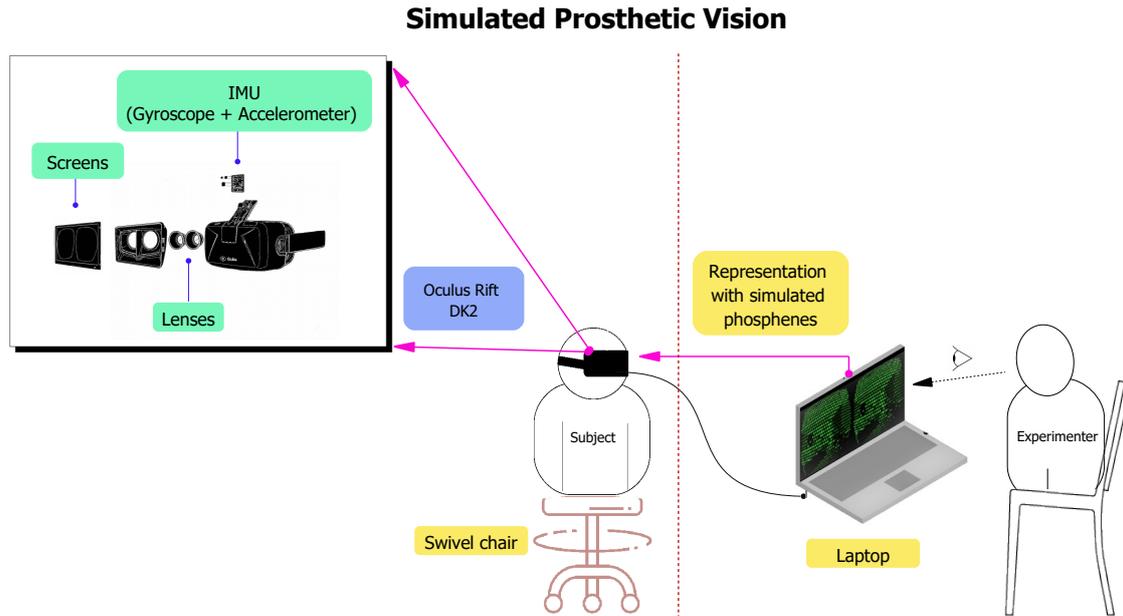


Figure 3.1: SPV system. The components consist of an Oculus Rift powered by a consumer level laptop. The VR system is composed by two lenses, two screens and a suite of internal sensors (gyroscope and accelerometer). The representation with simulated phosphenes is displayed on the laptop screen as well as on the Oculus system worn by the subjects. During the experiment, subjects were seated in a swivel chair allowing them to scan the entire scene all around them (360 degrees).

complex tasks because it supports realistic environments. This system acts as an electronic visual aid that attaches to the user’s head and presents information directly to the user’s eyes. The panoramic scenes (360 degrees) are intended to allow subjects to feel immersed in the scene.

3.2 Materials and Methods

We evaluate the influence of FOV on object search and recognition tasks using SPV through a VR system. The SPV system is a standard procedure for non-invasive evaluation using normal vision subjects. This methodology allows controlled evaluation of normally sighted subject response and task performance which is fundamental to know the way humans perceive and interpret phosphenized renderings. SPV also offers the advantage of adapting implant designs to improve the perceptual quality without involving implanted subjects.

3.2.1 Participants

Twenty four subjects aged 20-30 (6 female, 18 male) participated in the experiment. They had no visual problems or wore their normal optical correction during the experiment.

Ethics Statements

The research process was conducted according to the ethical recommendations of the Declaration of Helsinki. The research protocol used for this study is non-invasive, purely observational, with absolutely no-risk for any participant. There was no personal data collection or treatment and all subjects were volunteers. Subjects gave their informed written consent after explanation of the purpose of the study and possible consequences. The consent allowed the abandonment of the study at any time. All data were analyzed anonymously. The experiment was approved by the Aragon Autonomous Community Research Ethics Committee (CEICA).

3.2.2 Simulated Prosthetic Vision (SPV)

This section describes the SPV system including the hardware specifications, software components and phosphene generation.

Hardware

As shown in Figure 3.1, the experiment was conducted on an Oculus Rift powered by a consumer level laptop (Intel(R) Core(TM) i5-8265U CPU). This system is capable of working in real time with any mid-range laptop. The VR system, Oculus Rift DK2, is composed by two lenses, two screens and an inertial measurement unit (IMU) with gyroscopes and accelerometers, a standard setup for most commercial VR systems. It contains 5.7 inch dual OLED screens with a resolution of 960 x 1080 projected on to each eye. Each display is projected into the eye using a lens with pincushion distortion to provide peripheral vision. In our experiments we mostly use the central part of the display which remains undistorted. The representation with simulated phosphenes was displayed on the VR system worn by the participants as well as on the laptop screen for the experimenter to check the progress. For the head mounted display, we use a single channel (green) to avoid the chromatic aberration of the device lenses. During the

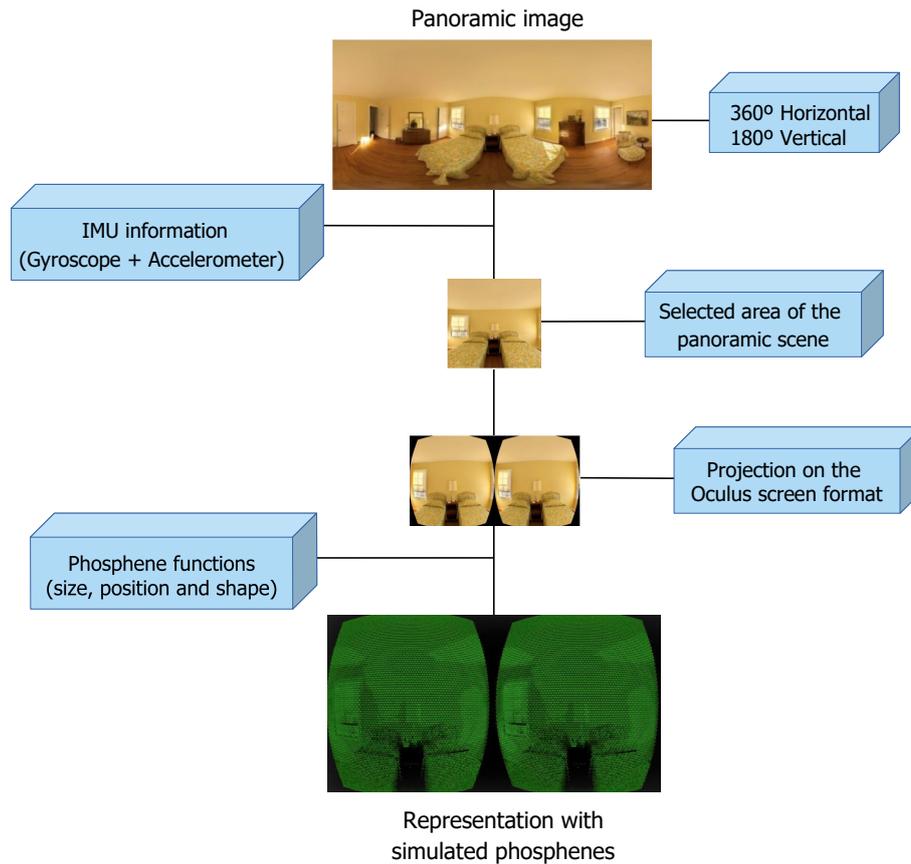


Figure 3.2: Data process. The input scene in our VR system is a panoramic image in equirectangular representation. The system estimates the head orientation using the IMU sensors (gyroscope and accelerometer) allowing to choose the area of the panoramic scene that is being observed at the moment. The area selected is then projected on the two Oculus screens and represented using simulated phosphenes.

experiment, participants were seated in a swivel chair allowing them to scan the entire scene with head rotation movements.

Software

The implementation was done in C++, using the Oculus Rift SDK to connect with the VR system and OpenCV for image processing. Our software is compatible with the Windows operating system. Figure 3.2 shows the data process designed to generate the stimuli for the VR system. Starting from a panoramic scene capturing 360° of horizontal FOV and 180° of vertical FOV, the system estimates the orientation of the user using the information collected from the IMU (Figure 3.1). The selected area is then projected on the two Oculus displays and finally represented with simulated phosphenes. The projection models for the VR system can be found in the Appendix.

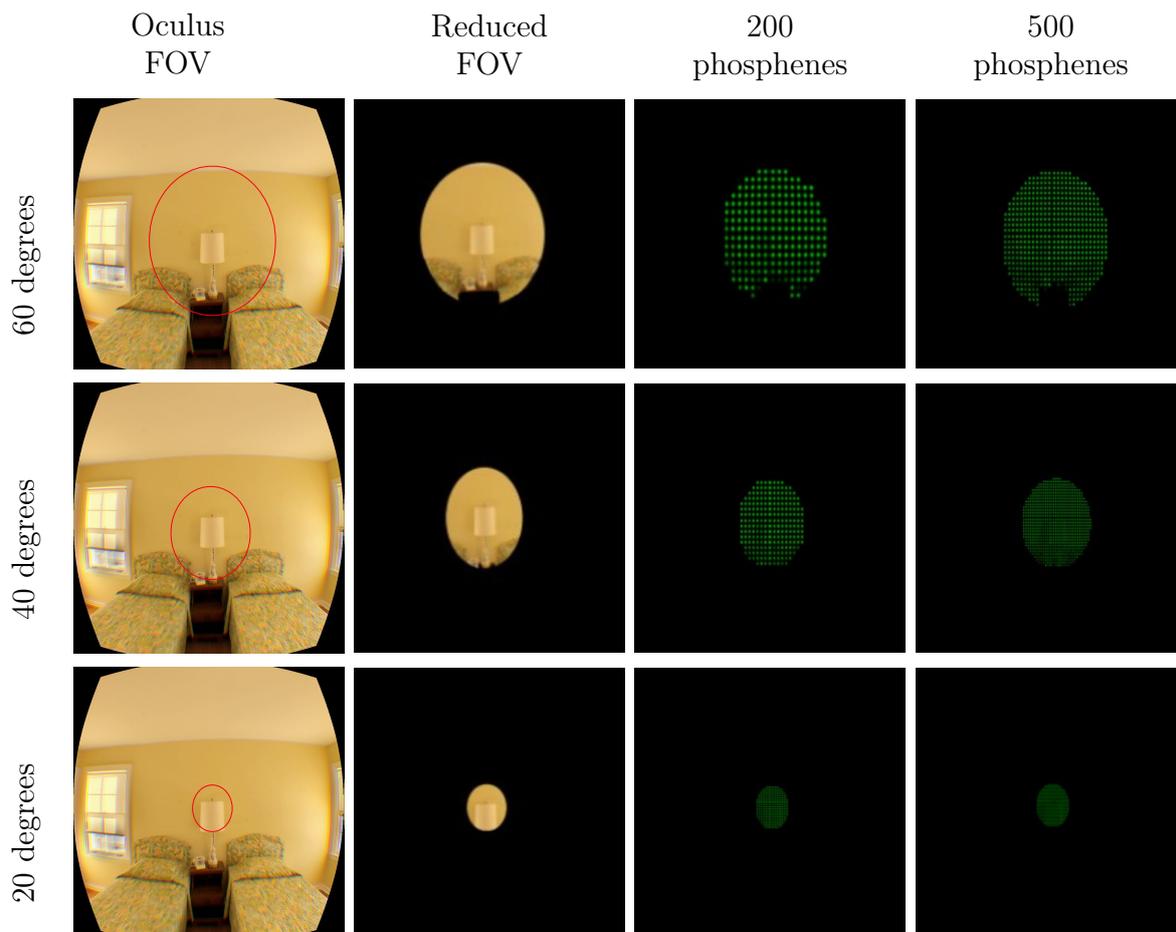


Figure 3.3: Stimuli conditions in the experiment. Rows: different FOVs used in the experiment (60, 40 and 20 degrees). Columns: different resolutions used in the experiment (200 and 500 phosphenes). Note that in the last row only part of the lamp is visible, the beds cannot be seen without moving the point of view.

Our phosphenes map configuration is similar to the frameworks of McKone et al. [48] and Chen et al. [45]. We approximate the phosphenes as circular dots with a Gaussian luminance profile –each phosphenes has maximum intensity at the center and gradually decays to the periphery, following a Gaussian function–. The intensity of a phosphenes is directly extracted from the intensity of the same region in the image. For our experiments, each phosphenes has 8 intensity levels. The size and brightness are directly proportional to the quantified sampled pixel intensities. The phosphenes map is calculated and updated with respect to head orientation in real time. The complete process of phosphenes generation can be found in the Appendix.

The software used in the experiment to simulate the prosthetic vision with the VR system is available at <http://webdiis.unizar.es/~rmcantin/index.php/Research/Vrfov>.



Figure 3.4: Object classes considered during the experiment. Subjects have to recognize the main objects of the scene, those objects that are usually present in hotel rooms such as bed, window, chair, tv, sofa, table/nightstand, door, wardrobe/shelving and lamp.

3.2.3 Procedure

The experiment was conducted using a selection of indoor panoramic scenes from a public database of Xiao et al. [178] that are adapted to our SPV system. The resolution of the panoramic scenes is 1024 x 512 pixels. All the scenes from the database are hotel rooms containing objects such as beds, tables, chairs, windows and doors, among others (see Figure 3.4). We removed several scenes because they had external distractions, such as signatures and watermarks, that could affect the experiment. From them, we randomly selected 50 scenes. The scenes were presented to the subjects using different stimuli conditions based on two resolutions (200 and 500 phosphenes) and three circular FOVs (20, 40 and 60 degrees), as can be seen in Figure 3.3. We used a circular FOV to avoid directionality in the searching process that might bias the results. We selected these particular resolutions and FOVs based on current retinal prostheses [32, 179], although our VR platform allows to quickly change those parameters. The total number of stimuli scenes generated for the experiment was 300.

For the formal experiment, participants were recruited to complete an object search and recognition task using the SPV system. By turning the swivel chair and head, subjects had to scan the entire scene by changing the head orientation, but not the position. At the same time, participants had to search and recognize the main objects of the scene such as *bed*, *window*, *chair*, *tv*, *sofa*, *table/nightstand*, *door*, *wardrobe/shelving* and *lamp* (see Figure 3.4). Small objects such as telephones, vases or remote controls were not taken into account. We took into account both the type of recognized object and the number of each of them. At the beginning of the experiment, subjects were trained during 2 minutes on a test set of images. Participants were informed that all scenes were indoor scenes,

but they were not informed about the class and number of objects in each scene nor the different stimuli conditions. The demo images were not included in the experiment to avoid learning effects.

Figure 3.5 shows the trial setup. Each trial consisted of a sequence of fifteen scenes generated by choosing a resolution, a FOV and a scene, among the shuffled conditions. Initially, a white dot was displayed in the center of the screen indicating where the participants had to maintain fixation until the beginning of the task. Next, each scene was presented and the user has a maximum of 60 seconds to scan the scene moving the chair and their head. The following scene was displayed when the participant verbally ordered that the task had been completed ('I cannot see more objects') or after 60 seconds. This procedure was repeated for each scenario of the trial. Between two scenes the participants were instructed to 'look for' the white dot to fixate the gaze in a steady position before the next scene. There was a break time in the middle of the sequence (between 7th – 8th scene) of 60 seconds.

The participants verbally indicated the type of objects seen in each scene during the scanning to minimize distractions. The responses of each stimulus scene were timed recorded and annotated by the experimenter. If the participants did not respond within the 60 seconds frame, the result of that scene was considered as 'no object was recognized' and a time of 60 seconds was recorded. Participants did not get feedback of their responses. The complete experiment took approximately 20 minutes per participant.

3.2.4 Statistical analysis

Data were analyzed using two-way ANOVA and post hoc-test with Tukey's method to evaluate simultaneously the effect of the two grouping variables (resolution and FOV) on the response variables object recognition and recognition time with $p = 0.05$, * < 0.05 , ** < 0.01 , *** < 0.001 and *ns* not significant.

3.3 Results

The performance for all resolutions and FOVs is summarized in Figure 3.6. The results show the ratio of recognized object and recognition time (mean \pm standard deviation) for aggregated data from all subjects and all images. For the same scene, we normalized

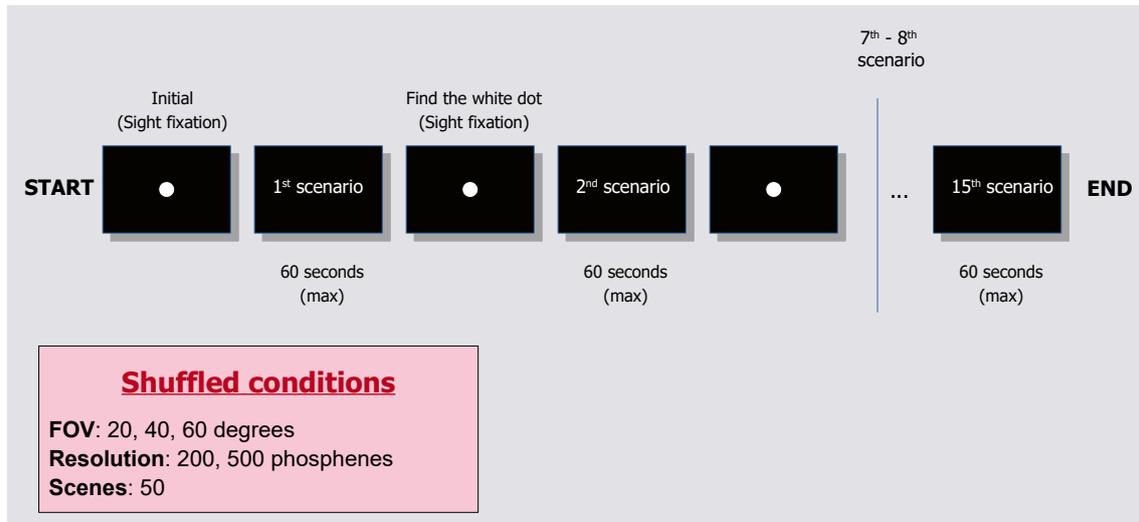


Figure 3.5: Trial setup. To generate the scene in each step of the trial sequence we used shuffled conditions of FOV, resolution and scenes. During the experiment, each scene appeared for 60 seconds and switched for the next scene automatically. Break time in the middle of the sequence (between 7th – 8th scene) was 60 seconds. The complete experiment took approximately 20 minutes.

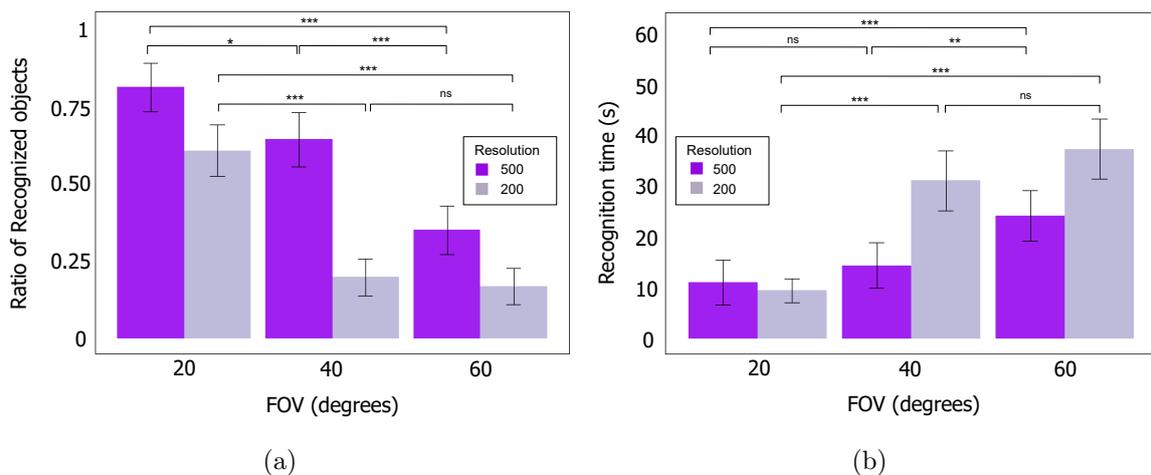


Figure 3.6: Ratio of recognized objects and recognition time. (a) Object recognition results for 20, 40 and 60 FOVs and for the two resolutions, 200 and 500 phosphenes. High scores indicate that subjects were able to recognize most of the objects in each scene. (b) Recognition time results for 20, 40 and 60 FOVs and for the two resolutions, 200 and 500 phosphenes. High scores indicate that the subjects needed more time to perform the recognition task. ***= $p < .001$; **= $p < .01$; *= $p < .05$; ns= $p > .05$.

the results of the object recognition with all the conditions of the experiment, with “1” being the experiment case with the highest number of recognized objects for a particular scene. The time recorded in each scene was the time it took the subject to recognize the first object. We also performed a test to determine if the mean difference between specific pairs of conditions are statistically significant using Tukey’s method with a significant

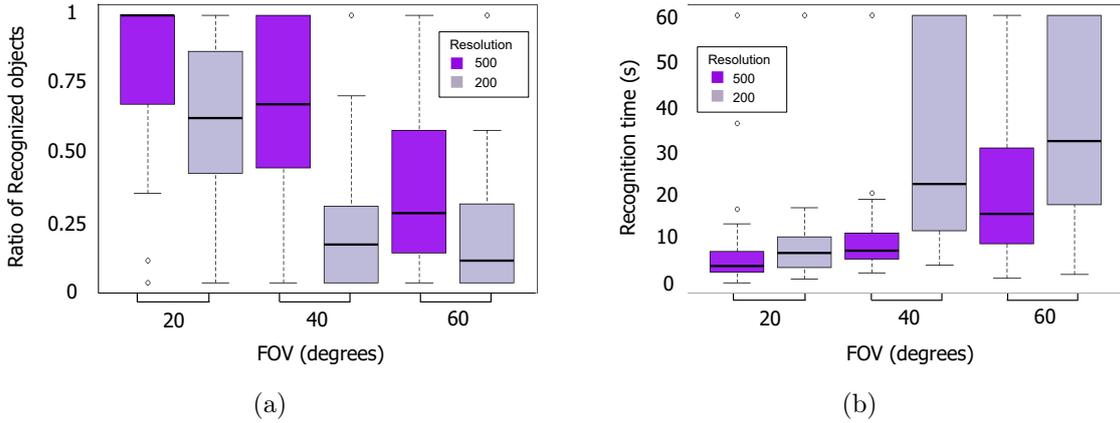


Figure 3.7: Ratio of recognized objects and recognition time. a) Box-plot for object recognition and b) box-plot for recognition time.

level $\alpha = 0.05$.

Figure 3.6(a) shows the object recognition performance for the three FOVs (20, 40 and 60 degrees) and the two resolutions (200 and 500 phosphenes). For the same resolution, the object recognition performance decreases as the FOV increases. For the resolution of 500 phosphenes the average performance is 80.66 ± 7.85 , 63.84 ± 8.68 and 34.74 ± 7.83 for 20, 40 and 60 degrees respectively. For the resolution of 200 phosphenes the average performance is 60.24 ± 8.28 , 19.65 ± 5.97 , 16.58 ± 5.82 for 20, 40 and 60 degrees respectively. No significant differences were found for 40-60 FOVs for 200 phosphenes ($p=0.8056$), as performance was very poor in both cases. Comparing the performance for the same FOV, the performance increases with the number of phosphenes for all cases.

Figure 3.6(b) shows the recognition time for the three FOVs and two resolutions. For the same resolution, the recognition time increases as the FOV increases. For the resolution of 500 phosphenes the average recognition time is $10.81 \pm 4.32s$, $14.10 \pm 4.41s$ and $23.71 \pm 4.91s$ for 20, 40 and 60 degrees respectively. For the resolution of 200 phosphenes the average recognition time is $9.24 \pm 2.35s$, $30.50 \pm 5.79s$, $36.61 \pm 5.77s$ for 20, 40 and 60 degrees respectively. No significant differences were found for 20-40 FOVs for 500 phosphenes ($p=0.733$) and similarly to 40-60 FOVs for 200 resolution ($p=0.199$). Comparing the performance for the same FOV, the recognition time decreases with the number of phosphenes. Table 5.1 shows the number of cases considered failure, where no object has been recognized in 60s. The condition 60 FOV and 200 resolution had more number of failures. Contrary, the condition 20 FOV and 200 resolution had fewer number of failures.

Figure 3.7 shows box-plots of data distribution with 25, 50 and 75th quartiles for recognized objects and recognition time. In the recognized object (see Figure 3.7(a)), there is difference between the two resolution (200 and 500 phosphenes) for 20 and 40 FOVs. On the other hand, for 200 phosphenes there is only difference between 20 and 40 FOVs. In the case of recognition time (see Figure 3.7(b)), there is difference between the two resolution for 40 and 60 FOVs. At the same time, there is difference between 40 and 60 FOVs for 500 phosphenes and between 20 and 40 FOVs for 200 phosphenes. Note that the data distribution for the condition 20 FOV and 200 resolution and the condition 40 FOV and 500 resolution are very similar, even though they are two totally different schemes. However, it makes sense because they have practically the same angular resolution (see Figure 3.8).

Figure 3.8 shows the ratio of recognized objects and time recognition according to *angular resolution* (AR), measured in phosphenes per radian. We perform a logarithmic regression for all data. The logarithmic regression equation for the *object recognition rate* (OR) is $OR = -1.0345 + 0.4482 \cdot \log(AR)$ with $R^2 = 0.4031$ and $F - value = 201.3$. The positive coefficient indicates that as the angular resolution increases, the object recognition ratio tends to increase (see Figure 3.8(a)). The logarithmic regression equation for the *recognition time* (RT) is $RT = 83.14 - 18.78 \cdot \log(AR)$ with $R^2 = 0.2344$ and $F - value = 91.26$. The negative coefficient indicates that as the angular resolution increases, the recognition time tends to decrease (see Figure 3.8(b)). The regression output shows that both, ratio of recognized object and the recognition time variables are statistically significant with $p - value < 0.05$ (see Table 5.1).

3.4 Discussion

Recent developments in the technology for retinal implants should allow to improve the quality of electrode arrays, in terms of features, such as response time, resolution or size [94, 173]. Thus, an open question for the prosthesis design is how to evaluate and predict the utility and functionality in terms of patient benefit with respect to design parameters [32]. Several studies have incorporated performance-based measures and questionnaires to understand the relation of visual parameters in the performance of everyday tasks in visually impaired subjects.

In this work, we study how certain features interact and affect perception. Intuitively,

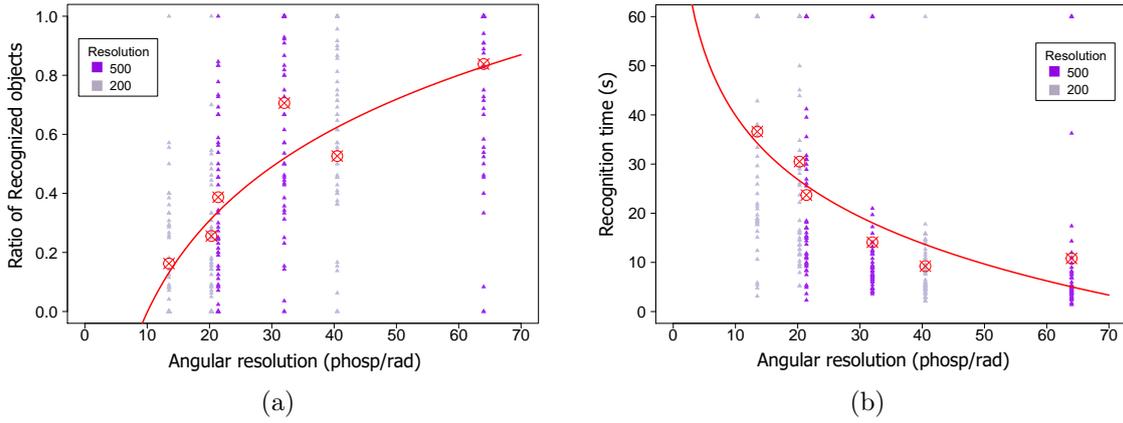


Figure 3.8: Angular resolution. (a) Performance of recognized objects according to angular resolution (phosphenes/radian) and (b) recognition time according to angular resolution. We perform a logarithmic regression for all data. The performance of recognized object tends to increase with angular resolution. Contrary, the recognition time tends to decrease with the angular resolution.

wider sensors should improve perception by allowing a larger field of view which can help detect motion and perform scanning [173]. Our results seem to indicate that it is not the case for head scanning, if it comes at the expense of phosphene density. Similarly, He et al. [176] investigated the effect of a wider FOV of a retinal prostheses on the performance of Argus II users in an object localization task using a thermal sensor. The results showed that users were able to find objects using the current $11^\circ \times 18^\circ$ FOV with no zoom with higher precision and speed than when using a wider FOV by zooming out the sensor input. They also suggest that a higher spatial resolution may be preferred over a wider input FOV. This has demonstrated the importance of some parameters such as FOV, visual resolution and angular resolution [180]. Others works have also shown that the lack of visual information caused by the low resolution and the restriction in a large portion of the FOV can be compensated by scene scanning [93, 181]. During scene scanning, the relevant information of the environment is actively sought, quickly and efficiently. For instance, there is some evidence on humans that eye and head rotation facilitates the learning process to recognize simple objects [181–184].

Simulated prosthetic vision (SPV) can be used to estimate the visual requirements to deliver a sufficient visual resolution and FOV with a large statistical power and cost efficiency. It provides an opportunity for simulation-based research regarding the design of a functional visual prosthesis, improvement of functional vision in low-phosphene-

Table 3.1: Mean value of the ratio of recognized object and recognition time for the six angular resolutions. Parameters for logarithmic regression and R^2 , F – value and p – value.

Angular resolution (phosp/rad)	Ratio of Recognized object	Recognition time (s)	# Results > 60s
13.5	0.166 ± 0.058	36.61 ± 5.77	20
20.3	0.197 ± 0.059	30.50 ± 5.79	14
21.4	0.346 ± 0.078	23.71 ± 4.91	7
32.0	0.638 ± 0.087	14.10 ± 4.41	4
40.5	0.602 ± 0.083	9.24 ± 2.35	1
64.0	0.807 ± 0.076	10.81 ± 4.32	4
<i>intercept</i>	-1.0345	83.14	–
<i>slope</i>	0.4482	-18.78	–
R^2	0.4031	0.2344	–
F – value	201.3	91.26	–
p – value	***	***	–

count devices, and also as a tool to search for image processing strategies to impart the most understandable prosthetic vision making experimentation with implanted patients unnecessary. In the work of He et al. [176], the FOV is modified at the external sensor by introducing a zoom lens, therefore changing the perceived scale. In contrast, thanks to the use of SPV we can actually change the properties of the electrode array, altering the FOV at the stimulus, maintaining the 1:1 scale ratio between perception and real world. Visual scanning has also been studied in SPV showing that it has a positive impact on the task performance by allowing subjects to increase the visual information [185]. Cai et al. [186] noticed during experiments that head rotation also allowed subjects to expand their effective FOVs and rudimentary depth perception through parallax. Chen et al. [187] also reported that maximizing visual information with scanning techniques improves visual acuity performance.

The theoretical resolution achievable by present-day retinal implants such as Argus II is 4° [188]. There have also been cases reported in Argus II clinical trials with 1.1° , well below the theoretical limit. This may be explained by effective scanning techniques, allowing subjects to temporally integrate percepts. Previous studies in SPV found out that a minimum angular resolution of 170 pixels/radian is needed for an acceptable accuracy in some tasks such as mobility and object recognition [38, 95, 177, 189]. However, this angular resolution is around three times more than the maximum angular resolution obtained in our experiment (64.0 phosphenes/radian).

We found that object recognition is well achieved with low resolution and restricted FOV. As can be seen in Figure 3.6(a), we obtained a significant improvement in the recognition performance as the FOV was reduced, for both resolutions. Besides, participants took less time to recognize objects with a narrower FOV (see Figure 3.6(b)). This seems counterintuitive since with a narrower FOV the global reference of the scene is lost. However, the narrower the FOV, the higher the angular resolution and therefore the greater the image detail (higher frequencies). Contrary, the widest FOV allows to cover the widest area of the scene but it only allows to see the gist of the image (lower frequencies).

Psychophysical and computational studies have shown that high and low spatial frequency provide different content from a scene: higher spatial frequencies contain fine information of image details and/or object boundaries, whereas lower frequencies preserve coarse blobs representing the gist of the scene [190–193]. In our experiments, the narrower FOV produces a higher phosphene density or angular resolution. By increasing the FOV (20, 40, 60 degrees), the angular resolution decreases: 40.5, 20.3 and 13.5 phosphenes/radian for 200 phosphenes and 64.0, 32.0 and 21.4 phosphenes/radian for 500 phosphenes, as can be seen in Figure 3.8. For lower angular resolutions, there were more cases of responses from subjects that exceeded the 60 s limit in the experiment, which implies that the subjects need more time to recognize details in these conditions (see Table 3.1). Our results suggest that it is better to see higher spatial frequencies of the scene, even if the global reference is lost due to the narrow FOV, since the subjects are able to holding back the global concept of the scene through visual scanning. Further, in more complex scenes such as low-contrast or low-luminance level, the time required to recognize the objects was higher and recognition performance decreased. In the same way, in those scenes with high contrast the subjects needed less time and the performance in the task increased. This fact has been demonstrated by Ehrlich et al. [184], where they observed that as the contrast increases the perception of the scene increases.

Virtual-reality (VR) systems have been widely used for experimentation [65, 69, 99, 194, 195]. They improve the quality of SPV tests since the experimentation environment is closer to the real-world. Some SPV research such as Denis et al. [194] used a VR system with two videos cameras mounted on the front of the headset to capture the visual scene for text localization. In [65], the SPV was generated on a VR headset with a stereo

camera that captured the scene in real-time for object recognition and localization. In similar studies, participants interacted with the VR system for different tasks such as visual acuity measurement [195] or wayfinding task [69]. Vergnieux et al. [99] used a virtual environment that was displayed via a VR system for navigation task evaluation.

Despite these proposals, we need more realistic and easier to use experimentation environments to improve the quality of SPV tests. A complete immersive experience implies to consider rotations and translations in the virtual environment. This approach would require the development of a detailed 3D model of the scene, to track the user and a large space allowing the free movement of the user in the real world. Since the influence of visual scanning in the FOV is dominated by rotations we have considered a simpler option. The environment is represented by a 360 panoramic scene so that the head can be turned in all directions from the center of the scene and the subject can explore the entire scene. This approach has a set of advantages. Panoramic scenes are easy to obtain in any real scene. By contrast, modelling a complete 3D model of the environment is very expensive. Further, the rotational movement of the head can be easily obtained from the embedded inertial sensors of the headset avoiding external cameras for the translation tracking. The experiment can be performed sitting on a chair, avoiding collisions, the requirement of a very large space and facilitating repeatability. Furthermore, our SPV system allows replication of the experiment using many subjects with normal vision without being limited by the number of implanted subjects.

3.5 Conclusions

We have analyzed the influence of field of view with respect to resolution in retinal prostheses through a study with a novel simulated prosthetic vision setup: a virtual-reality system using panoramic scenes. Participants perceived phosphene images in an immersive head-mounted device. The task consisted on finding and identifying common objects with different field of view and number of phosphenes. Our results show that, for the same number of phosphenes, recognition accuracy and response time improved by reducing the field of view. In fact, angular resolution is major determinant for effective object recognition, being directly correlated to the accuracy and inversely correlated to the response time. However, it has also shown a diminishing return even for an angular resolution of less than 2.3 phosphenes per degree. Simulated prosthetic vision allows

experiments with larger number of participants and simpler procedures than clinical studies with implanted patients. Our experimental setup relies on a consumer-level head-mounted display, public image databases and we have released the software needed to run the simulator, to facilitate replications and extensions. Our results indicate that the phosphene versus field-of-view trade-off for improvements to the angular resolution should prioritize the former to the latter.

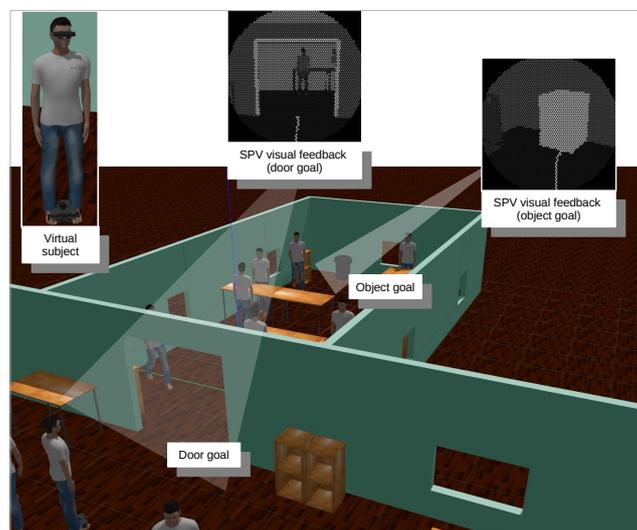
3.6 Related Publications

1. Sanchez-Garcia M., Martinez-Cantin R., Bermudez-Cameo J and Guerrero J. J. 2020. “Influence of field of view in visual prostheses design: Analysis with a VR system”. *Journal of Neural Engineering*

Chapter 4

4 Augmented reality navigation

The visual functions of visual prostheses seriously restrict the person's ability to navigate in unknown environments. Implanted patients still require constant assistance for navigating from one location to another. Hence, there is a need for a system that is able to assist them safely during their journey. In this work, we propose an augmented reality navigation system for visual prosthesis that incorporates a software of reactive navigation and path planning which guides the subject through convenient, obstacle-free route. It consists on four steps: locating the subject on a map, planning the subject trajectory, showing it to the subject and re-planning without obstacles. We have also designed a simulated prosthetic vision environment which allows us to systematically study navigation performance. Twelve subjects participated in the experiment. Subjects were guided by the augmented reality navigation system and their instruction was to navigate through different environments until they reached two goals, cross the door and find an object (bin), as fast and accurately as possible. Results show how our augmented navigation system help navigation performance by reducing the time and distance to reach the goals, even significantly reducing the number of obstacles collisions, compared to other baseline methods.



4.1 Introduction

Technological integration of Artificial Intelligence (AI) in the field of prosthetics and assistive technology has become a helpful tool for people with disabilities. Application of AI and robotics technology has a huge potential in achieving independent mobility and enhances the quality of life in persons with disabilities [196–200].

Augmented reality (AR) is showing day by day that has a place in our daily lives and can become a useful tool. Researchers have placed this new horizon to produce technology that is destined to change the way we solve problems and interact with the world. This ranges from work, education, medicine, and enhancing the independence of people with disabilities such as Retinitis Pigmentosa (RP) and Age-related Macular Degeneration (AMD) [77, 194, 201, 202].

RP and AMD are the two most prevalent retinal degenerative diseases. They affect millions of individuals worldwide and cause permanent blindness due to a gradual loss of photoreceptor [168, 203, 204]. Although current remedies can slow the progression of vision loss, there are no permanent cures for these retinal diseases [205, 206]. Retinal prostheses have turned out to a promising technology to improve vision in patients with RP and AMD. Visual prostheses can partially restore vision, bypassing damaged photoreceptors and electrically stimulating the surviving retinal cells, such as the retinal ganglion cells. For the blind, a retinal prosthesis can be life-changing, partially restoring sight, improving mobility and even helping to recognize some objects. The Argus II epiretinal prosthesis (Second Sight Medical Products Inc., Sylmar, CA) combines a miniature eye implant with a patient-worn camera and a processor to transform how the patient experience the world [25]. The camera acquires images from the real-world that are transmitted to a portable visual processing unit linked to the camera. The processed information is sent to the retina via electrical impulses in the implant by an electrode array. The stimulation can activate a group of neurons in a small localized area of the retina. Patients implanted with an array of electrodes in the visual system report the perception of reproducible white spots in the visual fields (called phosphenes) after stimulation. The brain interprets patterns of phosphenes in the restricted area as visual information. To date, the results of implanted subjects have shown to be promising, and all subjects have demonstrated improved function using the implanted prosthesis.

There are still physiological and technological limitations of the information received by implanted patients. Spatial resolution of prosthetic implants is limited by several factors, including electrode density, size, number and pitch, electrode contact, and visual encoding. Up to 600 or 1,000 pixels are required for restoration of useful vision; however, the majority of prostheses use 100 or fewer stimulating electrodes [207]. Besides, current systems provides a field of view of approximately $18^\circ \times 11^\circ$ in the retinal area, which correspond to the field of view covered by the electrode implant on the retina.

Because of the small number of implanted patients, it is difficult to improve prosthesis design through extensive clinical trials. To systematically progress on the design of retinal prostheses it is possible to rely on Simulated Prosthetic Vision systems (SPV). SPV opens the opportunity to evaluate potential and forthcoming functionality in early stages of design with not implanted subjects.

The SPV system usually consists in a computer screen for the presentation of static or dynamic phosphene images or in a camera mounted on a virtual-reality headset [73, 74, 77, 99, 131, 208, 209]. Participants perform various tasks perceiving a set of phosphenes mimicking the percepts elicited by a retinal prosthesis, while wearing the head-mounted display. For example, Fornos et al. [177] used SPV to study how the restrictions of the amount of visual information provided would affect performance on simple pointing and manipulation tasks. Hayes et al. [95] designed a set of tasks to assess performance of object recognition and manipulation and reading using different sizes of electrode array. In a posterior study, Dagnelie et al. [38] explored minimal visual resolution requirements of a simulated retinal electrode array for mobility in real and virtual environments, experienced by normally sighted subjects in video headsets. Sanchez-García et al. [78] evaluated the influence of field of view with respect to spatial resolution in visual prostheses using a virtual-reality environment system for more realistic SPV environments using panoramic scenes. However, they mainly focused on reading and object recognition.

More and more SPV navigation studies are being carried out [39, 99, 210]. Navigation is an important component of the self-reliance of the blind. The visual functions of patients are restricted enough to seriously affect the person's ability to navigate unknown environments. Some researchers have focused on mobility and obstacle avoidance. Cha et al. [143] investigated the feasibility of achieving visually-guided mobility without extra

information. They show that 625 phosphenes and a camera field of view of 30° were required to reach acceptable performances. More recent studies also focused on obstacle avoidance in a corridor [72, 75]. Other works showed that highlighting obstacles [211] or planar surfaces [212] improve the preferred walking speed [213]. Dagnelie et al. [38] explored minimal visual resolution requirements of a simulated retinal electrode array for mobility in real and virtual environments with a high contrast between the ground and the walls. Rheede et al. [69] also used a virtual environment to check if subjects were able to follow instructions and walk through a predetermined path. The results of Dagnelie et al. [38] and Rheede et al. [69] also demonstrated that the comprehension of the environment was not necessary for the subjects to follow a predetermined path. Successful localization is based on the perception of specific signals from the environment (landmarks), but also on the selection of an appropriate path [214]. This is precisely the trend that has been followed in Robotics in recent years [215–219]. The problem of navigation in Robotics has been extensively studied. We can define autonomous navigation as a set of methodologies that make possible to move a robot safely through the environment. Autonomous navigation and obstacle avoidance have been widely study in mobile robotics [220–223]. More concretely, the aim of navigation systems is to search an optimal or quasi-optimal path from the start point to the goal point with obstacle avoidance competence. The autonomous navigation system usually includes various tasks such as: planning, perception and control. Planning in mobile robots generally consists of establishing the mission, the route and the avoidance of obstacles, sometimes in the presence of uncertainty [224, 225]. In the case of robotics, autonomous navigation culminates by sending orders to the robot so that it moves in the calculated direction. In our case of people assistance, we look for a way to communicate the information obtained through a visual prosthesis.

In this work, we take advantage of the robotic algorithms, adapt them to human navigation, and include them in SPV using a virtual environment. Specifically, we propose an augmented reality navigation system for visual prostheses. This method consists on four tasks: locating the subject on a map, planning the subject trajectory, showing it to the subject and re-planning without obstacles. We also design a virtual environment which allows us to systematically study navigation performance based on prosthetic vision. The main issue is to determine whether the use of a smart prosthesis based on a robotic

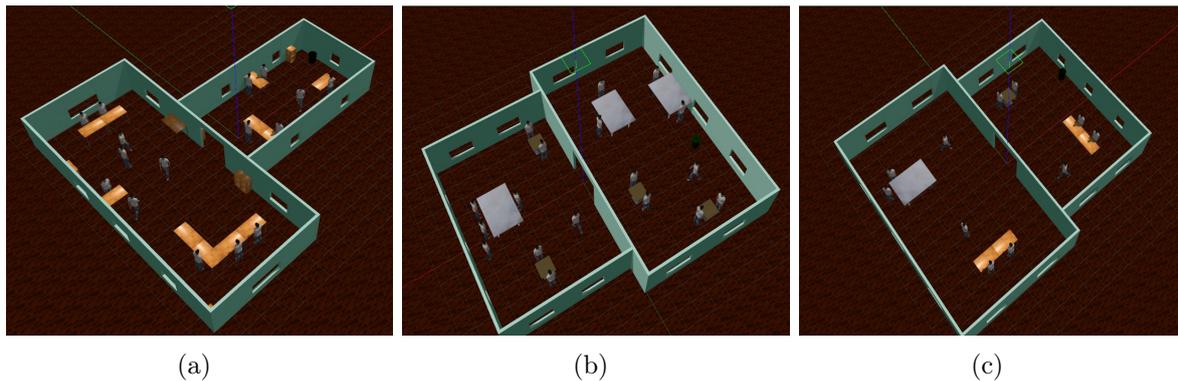


Figure 4.1: Virtual environments used in the experiments. (a) Environment 1, (b) Environment 2 and (c) Environment 3. Both, Environment 2 and Environment 3 have the same map but vary in the number and distribution of obstacles, being the Environment 2 more complicated than Environment 3, but less than Environment 1.

navigation system help in navigation tasks. We evaluate and compare the proposed guidance system with baseline methods through a SPV experiment, which is a standard procedure for non-invasive evaluation using normal vision subjects. The experiments included one task: finding and reaching a large object while avoiding obstacles in wildly crowded scenarios.

4.2 Methods

4.2.1 Subjects

Twelve subjects with normal vision volunteered for the formal experiment. The subjects (two females and ten males) were between 22 and 35 years old. Every subject used a computer daily (video games).

Ethical statement

The research process was conducted according to the ethical recommendations of the Declaration of Helsinki. The research protocol used for this study is non-invasive, purely observational, with absolutely no-risk for any participant. There was no personal data collection or treatment and all subjects were volunteers. Subjects gave their informed written consent after explanation of the purpose of the study and possible consequences. The consent allowed the abandonment of the study at any time. All data were analyzed anonymously. The experiment was approved by the Aragon Autonomous Community

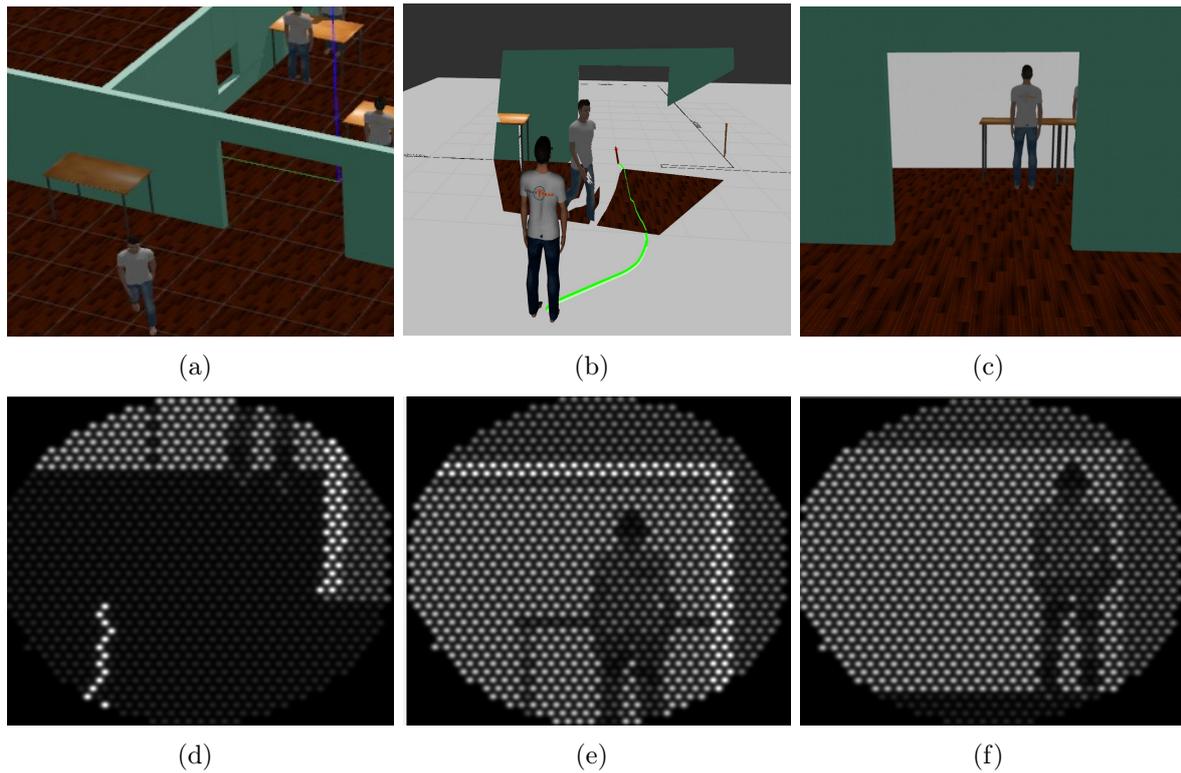


Figure 4.2: In this example, the subject’s goal is to cross the door. (a) Simulation environment showing the door goal. (b) The robotic system obtains an optimized trajectory to the goal avoiding obstacles (green line). (c) Image captured by RGB camera showing the door goal. (d) Image generated by the SPV with the RoboticG method. (e) Image generated by the SPV with the PerceptualG method. (f) Image generated by the SPV with the DirectG method.

Research Ethics Committee (CEICA, see Ethical Statement for additional details).

4.2.2 Augmented Reality System

The following subsections describe the three guidance methods used in the experiments. First, our proposal called Robotic Guidance (RoboticG) based on augmented reality navigation system highlights the optimized trajectory or path and the object goals on the phosphene representation. Second, we have used two guidance methods as baselines: a) we remove the path augmentation from our RoboticG method, leaving the goal perception (PerceptualG) and b) we remove both, the path and the goal augmentation from our RoboticG method (DirectG). Examples of the resulting effect on the phosphene images are shown in Figure 4.2 and Figure 4.3.

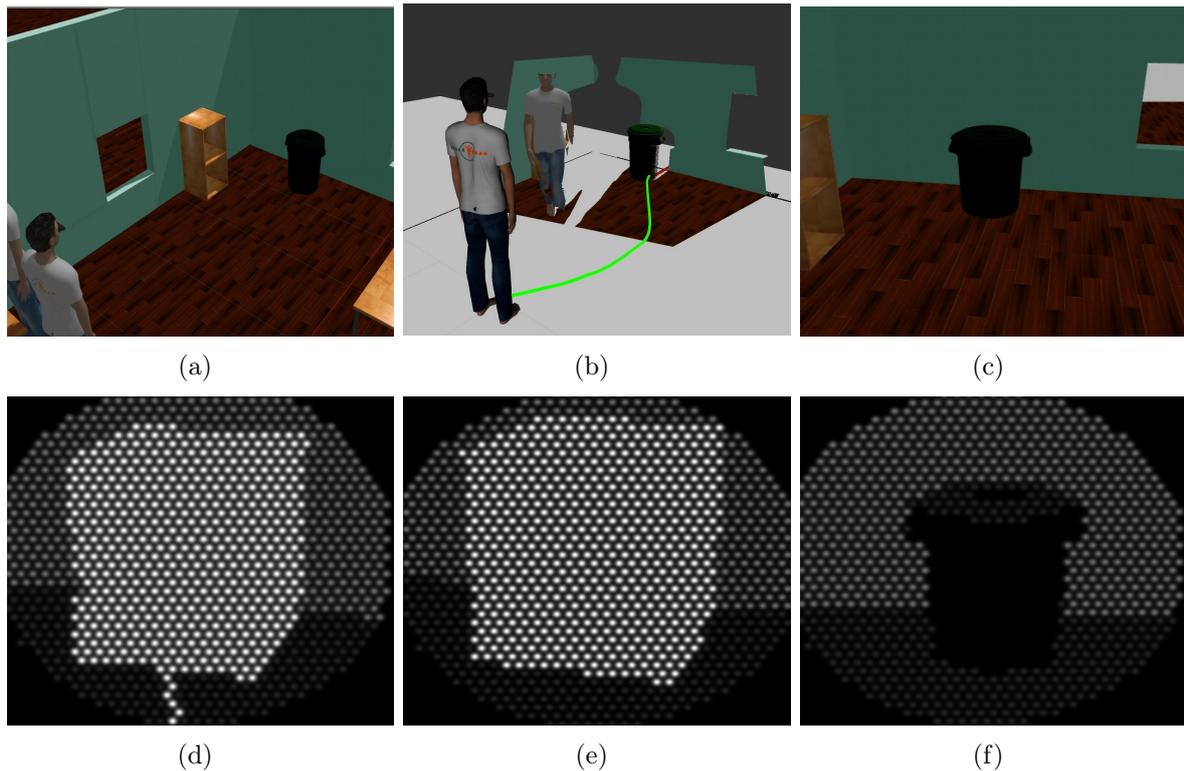


Figure 4.3: In this example, the subject’s goal is to find the bin. (a) Simulation environment showing the bin goal. (b) he robotic system obtains an optimized trajectory to the goal avoiding obstacles (green line). (c) Image captured by RGB camera. (d) Image generated by the SPV with the RoboticG method. (e) Image generated by the SPV with the PerceptualG method. (f) Image generated by the SPV with the DirectG method.

Robotic Guidance (RoboticG)

We propose an autonomous navigation system with obstacle avoidance for visual prosthesis called Robotic Guidance (RoboticG). This method performs a path planning that the person has to follow to reach a goal avoiding obstacles and highlight the path with phosphenes (augmented path), as can be seen in Figure 4.2(d) and Figure 4.3(d). Moreover, we use a goal perception method which segment the goals and highlight them with phosphenes (augmented goal), as can be seen in Figure 4.2(e) and Figure 4.3(e). We evaluate our RoboticG method using SPV. SPV is made up of Gazebo and simulated phosphenes.

Path planning The most important tasks for navigation are locating the user on the map and the path planning to reach a goal. For the location problem we determine the position and orientation of the subject relative to a 2D map of the virtual environment where it is going to be navigated. One of the great advantages of working in a simulated

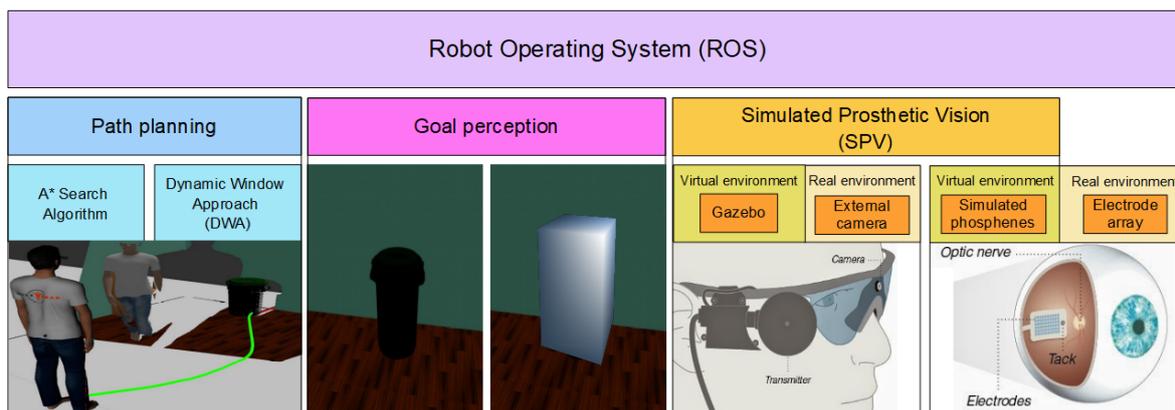


Figure 4.4: Diagram of SPV environment. We use the Robot Operating System (ROS) which handles communications between programs allowing easy control of a robot’s mobile operations. We evaluated our RoboticG method composed of Path planning and Goal perception using Simulated Prosthetic Vision (SPV). SPV is made up of Gazebo and simulated phosphenes. One advantage of our method is that in a real environment, Gazebo could easily be exchanged for the external camera of the prosthesis and the simulated phosphenes for the implant electrode array.

environment is that it allows us to know the user’s location on the map at all times. However, for real experiments, we could use a SLAM system [226] which allows locating the subject as long as the scene has previously been mapped. For both, the location system (3D map) and the navigation system (2.5D map), we use a common reference system that allows us to express information from the reference system of one map to the reference system of the other. Our simulation environment is precisely designed so that it can be used with a real sensor.

We use two planners for the path planning: a global planner and a local planner. The global planner is based on the A* Search Algorithm which calculates the optimal path between the source (initial state) and the destination (final state) using the 2D map (see Figure 4.4). The local planner is based on the Dynamic Window Approach (DWA) [227] which following the A* path, finds a collision-free (‘admissible’) trajectory considering the obstacles not included in the map. This algorithm uses the information of the obstacles from the 3D point cloud projected on a 2D map, to calculate the most optimal path with which to avoid them. The global trajectory is modified as a function of distance to which obstacles are located, the distance to the target and the alignment of the subject with respect to it. Obstacles are detected in real time.

Goal perception In order to augment the goals to highlight them, it is necessary to apply some kind of goal perception method, that allow us to recognize the goal element in the scene. Nevertheless, ‘goal perception’ is a broad term and the method to apply may need to be goal-specific. For instance, we could use the Region-Growing algorithm [228], deep learning detection/segmentation algorithms to identify certain objects in an image [120], or advanced 3D perception algorithms that leverage point clouds. Since object detection is not in the scope of this work, we leverage the known map and fix the goals (door and bin) to a certain point in the map so that they are unequivocally identified. This algorithm locates the objects/goals only when they are within a 5m radius.

Augmented representation with phosphenes We use a simple representation with phosphenes for our RoboticG method to represent the direction of the path and the goals. Since in prosthetic vision systems the spatial and photometric resolution is very low, we use a simple representation that is capable of communicating the clues for navigation without obstructing but rather complementing the useful information of the scene.

Our guidance method consists of representing a line on the floor plane (and projecting it on the image) that follows the path that the person must follow to reach a goal. To represent in the image the sequence of 2D points that leads to a goal calculated by the trajectory planner, we define the location of the goal on the 2D map and project the 3D points onto a phosphene mesh. We assign to these points a high intensity value to clearly differentiate them from the rest of the objects in the scene (augmented path). We also assign a high intensity value to the segmented goals by the goal perception method to differentiate them from the rest of the scene (augmented goal).

Simulated Prosthetic Vision (SPV) In this work, we simulated two components of the visual prosthesis: the external camera and the electrode array of the prosthesis, shown in orange in Figure 4.4. We simulate the external camera with the virtual environment called Gazebo, which is an open-source 3D robotics simulator [229]. The main advantage of Gazebo with respect to other simulators is that it is perfectly integrated with ROS, uses the same type of communication through topics and therefore facilitates a comfortable integration with the rest of the processes. Furthermore, sensor simulations produce messages of the same type as real sensors, so that the same code can be used

regardless of the source of the information. One of the main uses of Gazebo is the development of robotics algorithms in controlled simulation environments, so that they can then be applied to real systems without having to modify the algorithm itself. In this way, a person with a visual prosthesis could turn off the implant to stop seeing the real environment and plug in Gazebo seeing the 3D world in the physical implant. One advantage of our method is that in a real environment, Gazebo could easily be exchanged for the external camera of the prosthesis and simulated phosphenes for the implant electrode array.

In relation to the simulated electrode array, our simulated phosphene map is similar to the frameworks of Sanchez-Garcia et al. [77], McKone et al. [48] and Chen et al. [45]. We approximate the phosphenes as circular dots with a Gaussian luminance profile—each phosphene has maximum intensity at the center and gradually decays to the periphery, following a Gaussian function—. The intensity of a phosphene is directly extracted from the intensity of the same region in the image. For our experiments, each phosphene has eight intensity levels. The size and brightness are directly proportional to the quantified sampled pixel intensities. For our experiment, we use a field of view restricted to 20° and 1000 phosphenes. The augmented path and goals are highlighted with a higher phosphene brightness level when the subject is less than 5 m closer from the goal.

In this study, we have generated three virtual scenarios with Gazebo and use a dynamic human model in order to simulate the movements performed by a real person. With this two simulators we evaluate the operation of our RoboticG method in the fastest and most comfortable way as a previous step to taking it to a real environment, as can be seen in Figure 4.5.

Baseline

We perform an ablation study as used in AI by removing components from the RoboticG system to determine the significance of each component. First, we remove the path augmentation from our RoboticG method leaving the goal perception, as can be seen in Figure 4.3(e). This representation is called PerceptualG. Second, we remove any augmentation as can be seen in Figure 4.3(f). This method is called DirectG. Previous studies have shown that DirectG can be very effective in scenes where high contrast predominates [130].

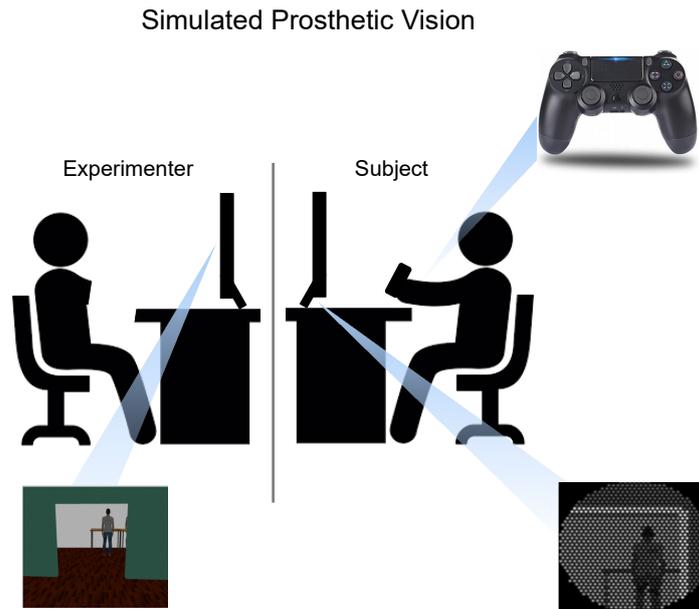


Figure 4.5: SPV system. Subjects were seated on a chair adjusted to their comfort in front of a computer screen holding a video game controller. The experimenter was in front of the subject seeing the same scene but in normal/RGB visualization.

4.2.3 Procedure

For the formal experiment, participants were recruited to complete a task: finding and reaching a large object while avoiding obstacles in wildly crowded scenarios. Each trial consisted of a sequence of scenarios presented randomly to the subject with the proposed RoboticG method, PerceptualG method and DirectG method, as can be seen in Figure 4.6.

Participants were seated on a chair adjusted to their comfort in front of a computer screen holding a video game controller, as can be seen in Figure 4.5. Participants were asked to navigate in a virtual environment while performing various tasks that involved locating an object and crossing the door of the room. Prior to starting the experiment, subjects were shown a training scenario with the normal/RGB visualization, as can be seen in Figure 4.2(c). The purpose of this training experiment was that subjects could become accustomed with controls, obstacles, goals and the experimental task. Participants were instructed to walk through the environment to familiarize themselves with the video game controller and the simulated environments. Participants also had to demonstrate to the experimenter that they were able to distinguish the goal from the obstacles by walking up to an example of both the door and the object. Following this introductory phase, subjects were given the training scenario with each of the navigation methods

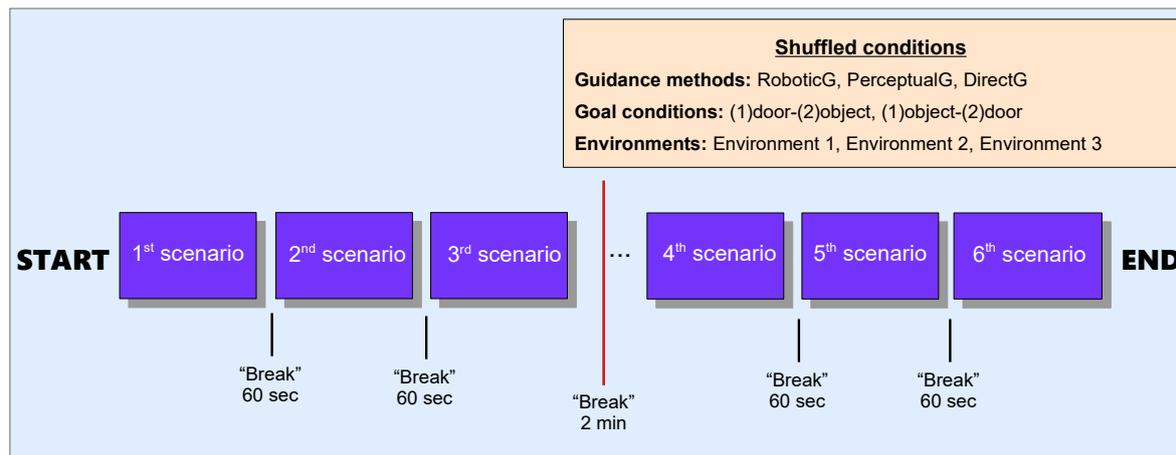


Figure 4.6: Trial setup. To generate the scenario in each step of the trial sequence we used shuffled conditions of environment, goal condition and guidance method. During the experiment, the break time after each scenario was 60 seconds. Break time in the middle of the sequence (between 3th–4th scenario) was 2 minutes. The complete experiment took approximately 30 minutes.

(RoboticG, PerceptualG and DirectG) in a random order and they were asked to perform the experimental task of finding the goals without colliding with any of the obstacles. Data from this session was recorded but not included in the analysis. Following the completion of the training trials the experiment was started.

There were twelve test conditions in total: 3 environments different from training (Environment 1, Environment 2 and Environment 3), 2 goals conditions ((1)door-(2)object: find the door first and then the object; or (1)object-(2)door: find the object first and then the door) and 3 guidance methods (RoboticG, PerceptualG and DirectG), as can be seen in Figure 4.6. Participants were instructed to perform six test scenarios with randomly chosen conditions. The participants were asked to walk through the scene until they find the door or object (depending on which goal the experimenter indicated to the subject to find first), avoiding the obstacles in the environment. Session duration was adjusted to allow completion of one trial given the participants' comfort and including short breaks every scenario. The break time after each scenario was 60 seconds. Break time in the middle of the sequence (between 3th–4th scenario) was 2 minutes. The complete experiment took approximately 30 minutes. Data acquisition for each participant was completed after six scenarios.

We gave additional aid to the participants if they were lost or expressed difficulties, e.g. 'take a few steps back' or 'turn right'. The number of additional aids was also registered.

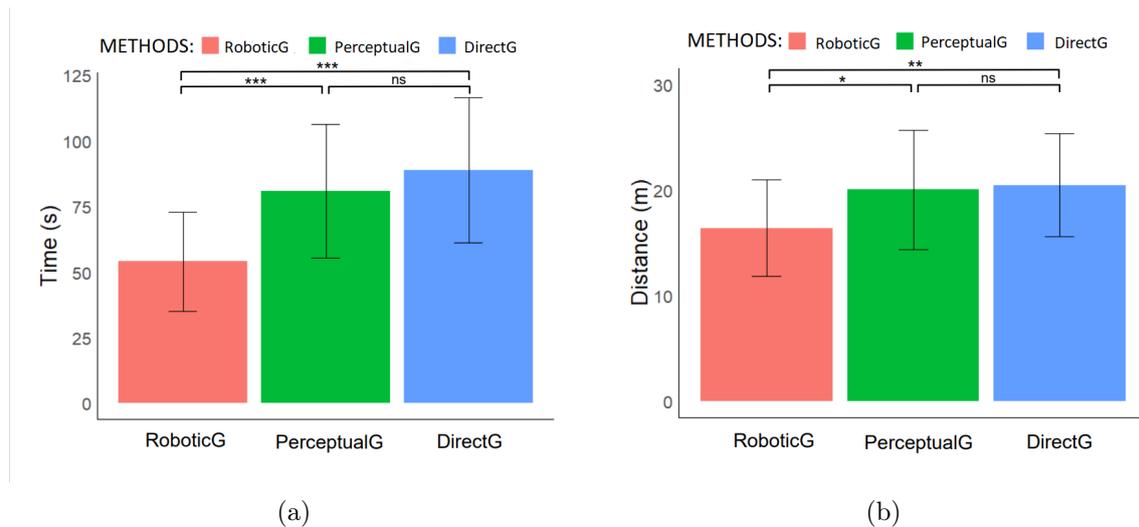


Figure 4.7: Mean time and distance required to reach both goals for the three guidance methods. (a) Time results for RoboticG, PerceptualG and DirectG methods. High scores indicate that the subjects needed more time to perform the location and avoiding obstacles task. (b) Distance results for RoboticG, PerceptualG and DirectG methods. High scores indicate that subjects covered a longer trajectory to reach the goals. ***= $p < .001$; **= $p < .01$; *= $p < .05$; ns= $p > .05$. All t-tests paired samples, two-tailed.

The performance was assessed according to the time elapsed from the beginning of the experiment until the subject found the second goal as well as the number of collisions. Subject positions were randomized for the three environments and for the goal order condition to decrease potential learning effects.

4.2.4 Statistical analysis

Data were analyzed using two-tailed paired t-test with Tukey’s correction to evaluate simultaneously the effect of the navigation methods (RoboticG, PerceptualG and DirectG) on the response variables covered distance and navigation time with $p = 0.05$, * < 0.05 , ** < 0.01 , *** < 0.001 and *ns* not significant.

4.3 Results

The performance for the three guidance methods is summarized in Figure 4.7. The results show the time and the distance (mean \pm standard deviation) for aggregated data from all subjects and all environments. Time (in seconds) is the time from the subject’s first step until the subject reach the goal. The distance (in meters) is defined as the covered distance since the subject start walking until the subject reach the goal. We also performed a test

Table 4.1: Mean value of time, distance and total number of bumps for the three methods (RoboticG, PerceptualG and DirectG). *** = $p < .001$; ** = $p < .01$; * = $p < .05$; $ns = p > .05$.

Guidance methods	Time (s)	Distance (m)	Total bumps	Helps by the experimenter
RoboticNG	54.02 ± 18.87	16.42 ± 4.59	4	0
PerceptualG	80.91 ± 25.47	20.03 ± 5.67	27	6
DirectG	88.80 ± 27.55	20.48 ± 4.85	24	9

to determine if the mean difference between specific conditions are statically significant using two-tailed test with a significance level $\alpha = 0.05$. The number of obstacle contacts or bumps within a trial was recorded as the number of times the distance between the center of an obstacle and the center of the subject's body is less than 1 m. The total number of bumps per method is shown in Table 5.1.

Figure 4.7(a) shows the time performance for the three guidance methods (RoboticG, PerceptualG and DirectG). For the DirectG method, the average performance is 88.80 ± 27.55 . For the PerceptualG the average performance is 80.91 ± 25.47 . For the RoboticG the average performance is 54.02 ± 18.87 . No significance difference were found for the DirectG and PerceptualG method ($p = 0.2512$), as time performance was very similar for both method. However, very significant differences were found for the RoboticG-PerceptualG and RoboticG-DirectG.

Figure 4.7(b) shows the covered distance for the three guidance methods. For the DirectG method, the average performance is 20.48 ± 4.85 . For the PerceptualG the average performance is 20.03 ± 5.67 . For the RoboticG the average performance is 16.42 ± 4.59 . No significance difference were found for the DirectG and PerceptualG method ($p = 0.7812$), as distance performance was very similar for both method.

Table 5.1 shows the total number of bumps for all the guidance methods. The PerceptualG method has more number of bumps than other methods. Interestingly, the methods DirectG and PerceptualG have almost the same number of collisions. However, it makes sense because in the PerceptualG method does not locate the goals until the subject is not close to them. This means that until reaching the goals the PerceptualG method is basically the same as the DirectG method.

Figure 4.8 shows box-plots of data distribution with 25, 50 and 75th quartiles for time and distance. In the case of required time to reach both goals (see Figure 4.8(a)),

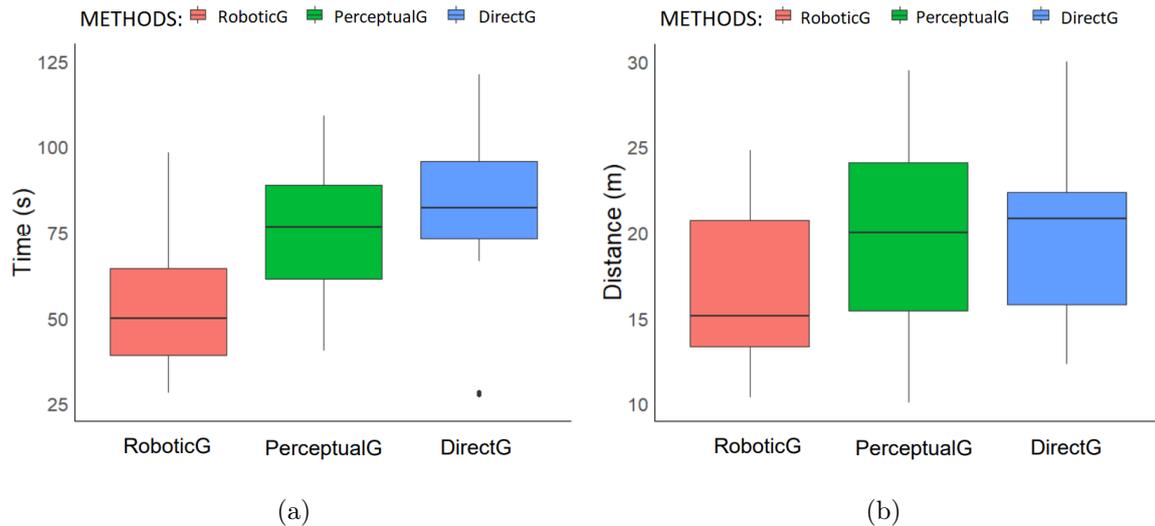


Figure 4.8: Mean time and distance required to reach both goals for the three guidance methods. (a) Box-plot for time. (b) Box-plot for distance.

there is difference between DirectG-RoboticG and PerceptualG-RoboticG. There is also difference between DirectG-PerceptualG. In Figure 4.8(b), the medians of the DirectG and PerceptualG methods are similar. However, comparing the medians of the DirectG and PerceptualG methods with RoboticG they are well separated, being the median for the RoboticG around 15m.

4.4 Discussion

Efficient navigation is a fundamental capability of every person to have autonomy on our daily life. People have their own well-defined criteria for choosing a path when moving from one place to another. But visually impaired people are not privileged to make their own choices when navigating. Significant amount of work has already been done for developing navigation systems for visually impaired people so that they could independently reach a target, without any kind of external assistance. Some examples of these systems are ultrasonic sensors, Global Positioning System (GPS) or radio-frequency identification [230, 231]. Very few researchers have focused on developing routing algorithms for the visually impaired [232–234].

In this work, we propose an augmented navigation system inspired on robotics research, called RoboticG, which combines a path planning algorithm and an obstacle avoidance method for routing the visually impaired person through an obstacle free optimal path.

There have been various methods introduced by researchers in different context to study path planning problem [235–237]. In the field of robotic, path planning is usually the first step to navigate a robot, where the final step incorporate techniques for obstacle avoidance [238–240].

We evaluated our RoboticG system using simulated prosthetic vision (SPV) composed by Gazebo [229], a virtual environment simulator, and the Robotic Operating System (ROS) [241]. Virtual environments can be built to implement the intended experimental platform without constraints associated with the real world. Furthermore, the use of virtual environments can help to avoid the experimental error caused by collisions and the use of touch to identify objects indirectly. Virtual environments have been widely used for experimentation in navigation tasks [38, 210, 242]. Some SPV research such as Vergnieux et al. [210] used a virtual indoor environment to investigate the navigation capabilities that could be restored through two different stimulation strategies consisting in a reduction of the environment view to match the number of electrodes and an object recognition algorithm in order to present only recognized elements. Wang et al. [39] created a virtual maze with SPV to assess navigation performance under contrasting conditions of varying luminance, background noise, and phosphene loss.

We found that object localization is well achieved with low resolution and restricted field of view using our guidance method (RoboticG). In a study by Golledge et al. [243], they conducted on the criteria of selection of roads or routes of an average human being and showed that the best classified criteria are the shortest distance, the least time and the least number of turns. As can be seen in Figure 4.8(a), we obtained a significant reduction in the navigation time until reaching the goals, compared to other baseline methods which do not used a navigation system, the PerceptualG and DirectG methods. Besides, participants took a shorter path with the RoboticG method compared with the baseline methods (see Figure 4.9). This is intuitive since the navigation algorithm RoboticG calculates the shortest path to reach the goals. Contrary, the participants took longer to reach the goals without the navigation system. Further, no significant difference was obtained with the baseline methods. This makes sense since the PerceptualG method locates the goals when the subject is at a distance less than 5m. From a real implementation, using the A* search algorithm we need to know where the goal is, but in practice the goal can be moved. Therefore, our navigation system gives an estimated

2D location of the goals. But when it comes to highlighting the goal we need a precise 3D location. In our case, the bin is placed on the ground but if the bin was not on the ground, we would need delicate perception. Therefore, our method leads the subject to an approximate location and identifies targets at 5 m. Until the subject is close enough to the targets, both the PerceptualG and DirectG methods are similar.

Obstacle avoidance also becomes a crucial point when dealing with visually impaired people. In any day-to-day environment, a visually impaired person may encounter various types of obstacles, whether dynamic or static. Early prototypes of obstacle detection devices were based on radar and sonar systems [244–246]. Recently, some of the available assistance systems focus on ringing a bell when an obstacle is detected on the road. However, these systems are unable to redirect the user from their current position to their desired destination once an obstacle has been detected. Our results show that the RoboticG method reduce drastically the number of bumps during the navigation due to the recalculation of the trajectory once an obstacle has been detected (see Table 5.1). One of the objectives of navigation systems is to calculate the optimal route that the user can take from his initial position to the destination. Figure 4.9 shows examples of path trajectories for the three guidance methods. In the three environments the path trajectories from the initial position to the goals were shorter and smoother with the RoboticG method. The PerceptualG and DirectG methods do not show much difference between them in environments 2 and 3. Although when we analyze it in a more complicated environment (with more obstacles) such as environment 1, we can see how an object recognize algorithm, PerceptualG, achieve a more direct path than the DirectG method which presents more oscillations. Moreover, our results suggest that subjects do not need additional aid with the navigation system from another person, in this case the experimenter (see Table 5.1). Although the RoboticG method has shown good results compared to the baseline methods, we believe that using only the augmented path of the RoboticG method (removing the augmented goal) would achieve very similar results.

This system has numerous tasks that have to be working simultaneously: point cloud processing, location on a map, path planning and generation of the phosphene image with the trajectory. All this is a challenge at the implementation level, since the numerous software packages involved have to communicate with each other and send the necessary information to each other on time. Our RoboticG implementation allows working in a

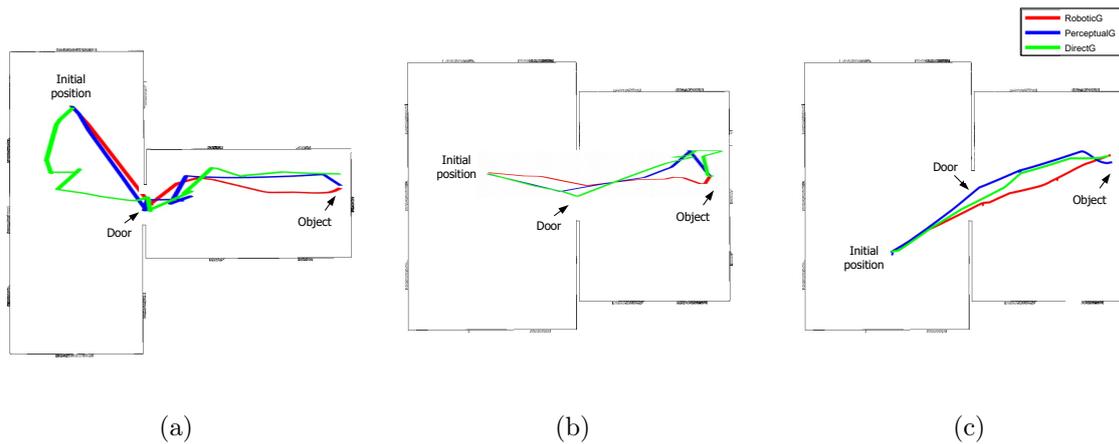


Figure 4.9: Examples of path trajectories for the three guidance methods. (a) Path trajectories in Environment 1. (b) Path trajectories in Environment 2. (c) Path trajectories in Environment 3.

simulation environment, which allows repeatability in the design of experiments with real subjects and gives statistical meaning to the results. Furthermore, the proposed guidance system works on ROS. ROS has been used to communicate and program the different systems of this project, a set of libraries and tools developed mostly in C++ focused on the development of robotic applications. One of the advantages of using a robotics framework such as ROS is that it allows us to extend the initial approach so that the system can be used in real scenarios using an RGB-D camera or using a SLAM system for localization [226].

4.5 Conclusions

We have proposed an augmented navigation system with obstacle avoidance for guidance in visual prosthesis which point out the path that the subject have to follow to reach the goals. Our motivation is that navigation in day-to-day environments is fundamental for all humans. Implanted patients still require constant assistance for navigating from one location to another. Hence there is a need for a system that is able to assist them safely during their journey. By using a route planning algorithm, the system route the subject through a shorter, obstacle-free route. We also designed a simulated prosthetic vision environment which allowed us to systematically study navigation performance. Twelve subjects participated in the experiment. The experiments included two tasks: object localization and obstacle detection and avoidance. Subjects were guided by the

visual guidance produced by the navigation system and their instruction was to navigate through different virtual environments until they reached a goals. Results show how our autonomous navigation system help navigation performance by reducing the time and distance to reach the goals, even significantly reducing the number of obstacles collisions, compared to other baseline methods.

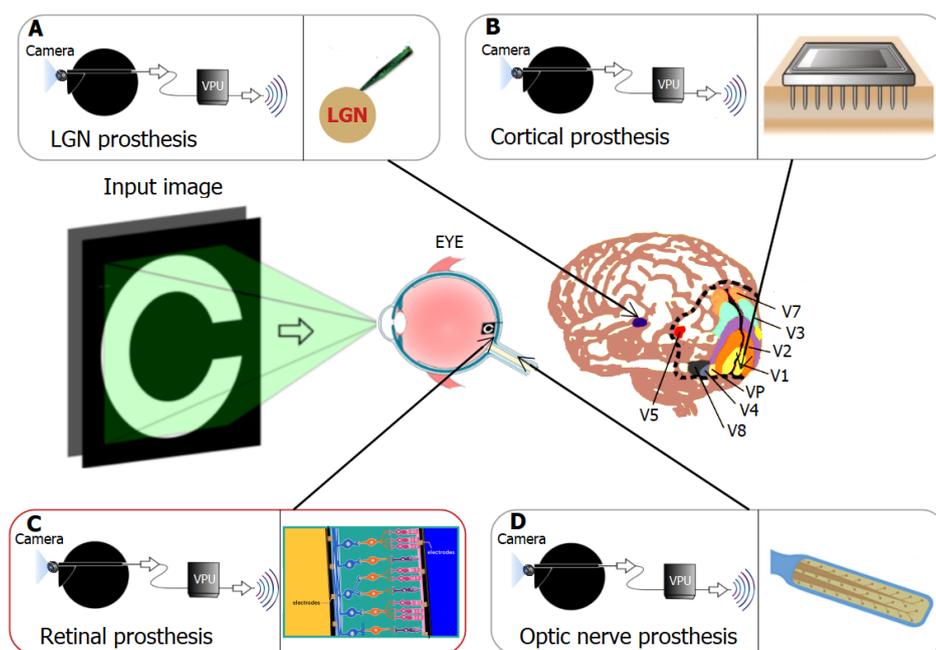
4.6 Related Publications

1. Sanchez-Garcia M., Perez-Yus A., Martinez-Cantin R., and Guerrero J. J. 2021. “Augmented reality navigation system for visual prosthesis”. Submitted.

Chapter 5

5 Visual acuity assessment with VR

In this chapter, we took advantage of VR software paired with a portable head-mounted display and evaluated the performance of normally sighted participants under SPV with variable field of view and number of pixels. In our system, the head-mounted display mimics the external camera of a retinal implant subjects and allows simple experimentation in order to study the design parameters of future visual prostheses. Subjects were required to identify different stimulus based on light perception, time-resolution, light location, motion perception and Landolt-C orientation commonly used for visual acuity examination in the sighted. Our results showed that of all conditions, a FOV of 20° and 1000 phosphenes proved optimal, with a visual acuity of 1.3 logMAR. Furthermore, performance appears to be correlated with phosphene density, but showing a diminishing return when FOV is less than 20° .



5.1 Introduction

Low vision or blindness are major health issues for the individual's quality of life. The leading causes of blindness are primarily age-related eye diseases such as age-related macular degeneration (AMD) [247, 248], cataract, retinitis pigmentosa (RP)[79] and glaucoma [249]. The loss of photoreceptors due to degeneration is a major cause of vision loss, resulting in dysfunctional light detection, transduction, and transmission [168, 203, 204]. In the case of RP, these inherited disorders can affect either rods or cone primarily. The most common form of RP is characterized initially by night blindness, followed by progressive loss in the peripheral field of view in daylight, eventually leading to blindness after several decades. The case of AMD is characterized by sudden acuity loss. Currently, there is no cure for RP or AMD [205, 206]. Nevertheless, different ways of restoring the resulting poor visual function have been studied, relying mainly on gene therapy, stem cell transplantation or visual prosthesis [32, 250–254].

Visual prostheses are presently the most viable technology for the treatment of low vision and there are many types being proposed because of its potential for the development of various types of devices with existing technologies. The basic concept of a visual prosthesis is 'electrically stimulating nerve tissues associated with vision (such as the retina) to help transmit electrical signals with visual information to the brain' [255]. Thus, various groups in the field are focusing on the development of new approaches for artificial vision based on electric stimulation of the retina [33, 157, 256], optic nerve [257–259], lateral geniculate nucleus [37, 260], or visual cortex [261–265]. All of these prosthetic devices work by exchanging information between the electronic devices and different types of neurons, and although most of them are still in development, they show promise of restoring vision in many forms of blindness.

At present, retinal prostheses are the most successful approach in this field [33, 157]. In a retinal implant a multielectrode array is set up on the retinal surface and stimulates the retina from the top side with electrodes. All this current implantable system are designed with electrodes implanted in the body working together with several devices worn outside the body. Thus, a visual prosthesis incorporates an external video camera for image acquisition, an image processor converting the image to a suitable pattern of electrical stimulation, and finally the electrical stimulation array on the retina itself [88, 266–268].

In spite of reports showing retinal prostheses capable of helping some participants perform simple tasks of daily living, such as detecting lights, recognizing objects and even reading large letters, there are still physiological and technological limitations of the information received by implanted patients. The number of electrodes and implant size limit the maximum amount of information that can be provided by the stimulating array. This fact has restricted the degree of visual resolution (up to 1500 phosphenes) and dynamic range of the visual perception (8 grey levels) that can be delivered to the user. Besides, current systems such as retinal implants provide a field of view (FOV) of approximately $18^\circ \times 11^\circ$ in the retinal area, which correspond to the FOV subtended by the electrode implant on the retina. Moreover, the visual acuity of existing devices is very low, which means that crucial skills such as facial recognition or navigation in unknown environments are not yet possible.

The visual acuity of prosthetic vision is limited by various factors from both engineering and physiological perspectives [269]. One of the main causes of low visual acuity is the limited spatial resolution that can be achieved by electrical stimulation with existing retinal implants. The size of the electrodes in today's retinal implants is often much larger than the size of the neurons in the retina, and the number of electrodes is low [270]. In addition, it has been shown that increased FOV is associated with a significant improvement in visual acuity [173]. Current retinal prostheses already approved for commercial use such as the Argus II epiretinal implant (Second Sight Medical Products Inc.) [156, 157, 271] and the Alpha-IMS subretinal implant (Retina Implant AG) [30, 272] have yet to provide sufficient visual acuity to allow blind patients to lead independent lives. These devices have been shown to restore vision up to a visual acuity of 1.8 logMAR and 1.44 logMAR, respectively. Regardless, the optimal number of the electrodes and FOV to provide adequate prosthetic vision is still an open question and an important design parameter needed to develop better implants.

In order to gain a better understanding of the potential benefits of low resolution visual prostheses we can use Simulated Prosthetic Vision (SPV). The SPV system is a standard procedure for non-invasive evaluation using normal vision subjects. Generally, in SPV, low-resolution image of the view is presented on a computer screen [77] or on a head-mounted display [78, 195, 273] glasses to a normally sighted user (see Figure 5.1). It allows researchers to rigorously investigate the minimal requirements for a functional

visual prosthesis and to explore which variables are important in the development of a visual prosthesis. Furthermore, prosthetic vision can be assessed in controlled, real, or virtual environments. Modern commercial virtual-reality (VR) systems have specifications in terms of resolution, latency and response time that allows a fully immersive experience with inexpensive equipment and computers. Using these devices, subjects are immersed and able to interact with complex environments.

Some researchers have used the combination of SPV with a VR system for experimentation. Sanchez et al. [78] analyzed the influence of field of view with respect to resolution in visual prostheses through a study with a SPV setup using a VR system. Kasowski et al. [40] proposed to embed biologically realistic models of SPV in immersive VR so that sighted subjects can act as virtual patients in real-world tasks. Thorn et al. [274] implemented prosthetic vision in a VR environment in order to simulate the real-life experience of using a retinal prosthesis and investigated the interaction between the field of view and the pixel number. Visual acuity has also been tested under a VR simulation for rectangular and hexagonal phosphene grids [195]. Chen et al. [273] examined visual acuity of prosthetic vision under VR simulation measuring parameters such as filtering scheme, filter aperture and the phosphene matrix.

In this work, we assess visual acuity in visual prostheses. We took advantage of virtual-reality software paired with a portable head-mounted display and evaluated the performance of normally sighted participants under simulated prosthetic vision with variable field of view and number of pixels. Our simulated prosthetic vision system allows simple experimentation in order to study the design parameters of future visual prostheses. The VR system acts as an electronic visual aid that attach to the user's head and presents information directly to the user's eyes.

5.2 **Methods**

We examined visual acuity on a stimuli recognition task using SPV through a VR system. The SPV system is a standard procedure for non-invasive evaluation using normal vision subjects. This methodology allows controlled evaluation of normally sighted subject response and task performance which is fundamental to know the way humans perceive and interpret phosphened renderings. SPV also offers the advantage of adapting implant designs to improve the perceptual quality without involving implanted subjects.

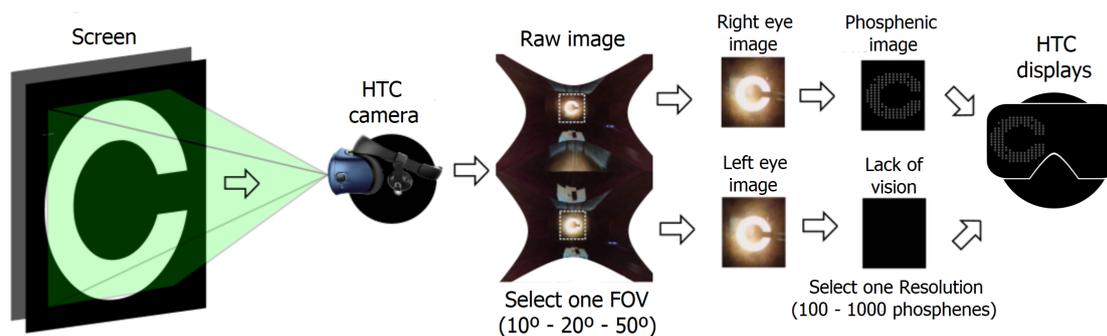


Figure 5.1: Data process. From the cameras of the VR HTC Vice Pro, we capture the raw images of the computer screen for the left and right eye. We select the central segments of the image to avoid distorted edges produced by the cameras and convert it to phosphenes. We simulate a single eye on, in this case the right eye and we turn off the left eye.

5.2.1 Participants

Ten subjects with normal vision volunteered for the formal experiment. The subjects (four females and six males) were between 22 and 35 years old. Every subject used a computer daily.

Ethical statement

The research process was conducted according to the ethical recommendations of the Declaration of Helsinki. The research protocol used for this study is non-invasive, purely observational, with absolutely no-risk for any participant. There was no personal data collection or treatment and all subjects were volunteers. Subjects gave their informed written consent after explanation of the purpose of the study and possible consequences. The consent allowed the abandonment of the study at any time. All data were analyzed anonymously. The experiment was approved by the Aragon Autonomous Community Research Ethics Committee (CEICA, see Ethical Statement for additional details).

5.2.2 Simulated Prosthetic Vision (SPV)

This section describes the SPV system including the hardware specifications, software components and phosphene generation.

Hardware

The experiment was conducted on an HTC VIVE PRO powered by a computer (Intel(R) Core(TM) i9-9900KF CPU 3.60GHz, NVIDIA GeForce RTX 2080 Ti). The VR system is composed by two lenses, two screens, SteamVR Tracking, G-sensor, gyroscope, proximity and Eye Comfort Setting (IPD). It contains dual AMOLED 3.5" diagonal screen with a resolution of 1440 x 1600 pixels per eye (2880 x 1600 pixels combined), covering a visual field of approximately 110 degrees. In our experiments we mostly use the central part of the display which remains undistorted. The representation with simulated phosphenes was displayed on the VR system worn by the participants as well as on the computer screen for the experimenter to check the progress. During the experiment, participants were seated in a backless chair allowing them to scan the entire scene with head rotation movements.

Software

The implementation was done in C++, using OpenVR for HTC VIVE Pro to connect with the VR system and OpenCV for image processing. Figure 5.1 shows the data process designed to generate the stimuli for the VR system. We watch the camera FOV to the viewfinder FOV. The selected area is projected on the two HTC displays. Finally, we convert the images into simulated phosphenes. We only project the phosphenic image to the right eye, simulating an implant in the right eye, and lack of vision on the left (see Figure 5.1).

Our phosphene map configuration is similar to the framework of Sanchez et al. [78]. We approximate the phosphenes as circular dots with a Gaussian luminance profile —each phosphene has maximum intensity at the center and gradually decays to the periphery, following a Gaussian function—. The intensity of a phosphene is directly extracted from the intensity of the same region in the image. For our experiments, each phosphene has eight intensity levels. The size and brightness are directly proportional to the quantified sampled pixel intensities. The phosphene map is calculated and updated with respect to head orientation in real time.

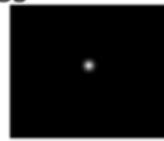
FOV \ Resol.	10°	20°	50°
100 phosphenes	C1 	C3 	C5 
1000 phosphenes	C2 	C4 	C6 

Figure 5.2: The six possible stimulus conditions are depicted for the ‘Landolt-C orientation test’. ‘FOV-Resolution’: C1: 10°-100 phosphenes, C2: 10°-1000 phosphenes, C3: 20°-100 phosphenes, C4: 20°-1000 phosphenes, C5: 50°-100 phosphenes and C6: 50°-1000 phosphenes.

5.2.3 Procedure

The experiment was conducted using a selection of stimulus from Balm [275] and Freiburg [276] tests, adapted to our SPV system, which are automated procedures for self-administered measurement of visual acuity. The images were presented to the subjects using different stimuli conditions based on two resolutions (100 and 1000 phosphenes) and three FOVs (10, 20 and 50 degrees), as can be seen in Figure 5.2. We selected these particular resolutions and FOVs based on current visual prostheses [30, 32, 156, 157, 271, 272], although our VR platform allows to quickly change those parameters.

The participants performed five tests based on different types of stimuli, described as ‘light perception’, ‘time recognition’, ‘light location’, ‘motion perception’ and ‘Landolt-C orientation’ (see Figure 5.3). The first stimulus corresponded to the ‘light perception’ and is the simplest stimulus of the experiment that tests the basic perception of light. The subjects’ task was to decide whether they see the light appear after the warning tone or not. The second stimulus corresponded to ‘time resolution’, which assesses one basic aspect of time resolution—namely, whether one or two flashes occur after an indicator beep. The third stimulus corresponded to the ‘light location’ that tests the projection of light. A light disc appeared that the subject must center in the limited visual field. After a pre-set delay, simultaneously with a warning tone, a wedge appeared directed up, down, right, or left from the fixation disc. The fourth stimulus corresponded to ‘motion perception’. A

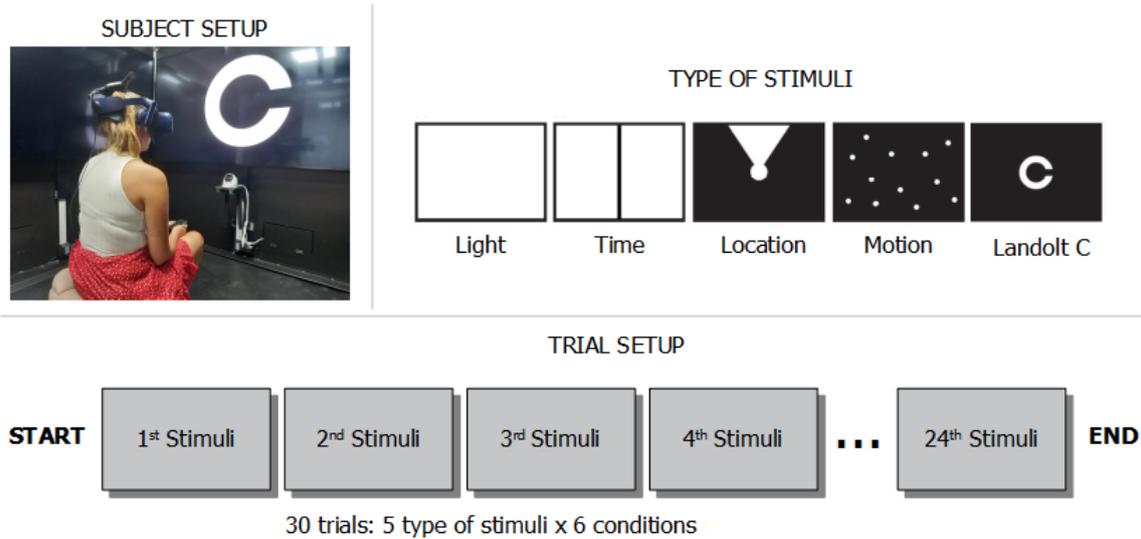


Figure 5.3: Subject and trial setup. Subjects view through the head-mounted display what is shown in the picture on the monitor. Subjects scanned the underlying picture with their head motion. They used the joystick to indicate the orientation of the stimuli for subject response. For the experiment, we used five stimuli from Balm [275] and Freiburg [276] tests: Light, Time, Location, Motion, and Landolt C. Each test consisted of 24 stimuli. The stimuli in each test were randomly selected.

random hexagonal pattern of light and dark elements appeared. After an acoustic signal, it began to move in one of four directions (up-down-right-left). The subject indicated the motion's direction. The last stimulus corresponded to the 'Landolt-C orientation' test. Subjects had to indicate the orientation in one of four directions (up-down-right-left) of the gap in the Landolt-C, which is a standard international symbol for testing visual acuity. In all tests subjects responded via corresponding joystick positions. The number of trials was set to 24 for all the tests.

5.2.4 Statistical analysis

Data were analyzed using two-way ANOVA and post hoc-test with Tukey's method to evaluate simultaneously the effect of the two grouping variables (resolution and FOV) on the response variables performance and reaction time with $p = 0.05$, $* < 0.05$, $** < 0.01$, $*** < 0.001$ and *ns* not significant.

5.3 Results

The results are summarized from Figure 5.4 to Figure 5.8. The results show the performance and the reaction time (mean \pm standard deviation) for aggregated data from all subjects. The performance (in percentage) is defined as number of correct responses. Time (in seconds) is the time from the subject's first response. We also performed a test to determine if the mean difference between specific pairs of conditions are statistically significant using Tukey's method with a significant level $\alpha = 0.05$.

5.3.1 Light perception

Figure 5.4 shows the performance and reaction time for the 'light perception' test. This test is the simplest of the five tests carried out in the experiment. All subjects were able to perceive light or not in almost all trials. All the conditions obtain high values above 95%, as can be seen in Figure 5.4(a). There was no significant difference between the stimulus conditions. The reaction time for all conditions are similar and close to 1.6s (see Figure 5.4(b)). The lowest reaction time was $1.47 \pm 0.34s$, corresponding to the condition of 20° and 1000 phosphenes. For the resolution of 100, no significant difference was found for 10-20 FOVs ($p=0.9964$), 20-50 FOVs ($p=0.7723$) and 10-50 FOVs ($p=0.8174$). For the resolution of 1000, nor significant difference was found for 10-20 FOVs ($p=0.2497$), 20-50 FOVs ($p=0.7680$) and 10-50 FOVs ($p=0.6223$). There was no significant difference between the two resolutions ($p > 0.05$). Some outliers can be observed in both graphs corresponding to instants in which the subjects were distracted during the task.

5.3.2 Time resolution

Figure 5.5 shows the performance and reaction time for the 'time resolution' test. For the resolution of 100 phosphenes, the average performance is 96.15 ± 4.53 , 96.70 ± 4.64 and 79.55 ± 11.78 for 10, 20 and 50 degrees respectively (see Figure 5.5(a)). No significant difference was found for 10-20 FOVs ($p=0.9863$). However, significant difference was found for 10-50 and 20-50 FOVs. For the resolution of 1000, the average performance is 95.05 ± 4.82 , 98.35 ± 2.66 and 97.80 ± 2.84 for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.1163$), 20-50 FOVs ($p=0.2160$) and 10-50 FOVs ($p=0.9369$).

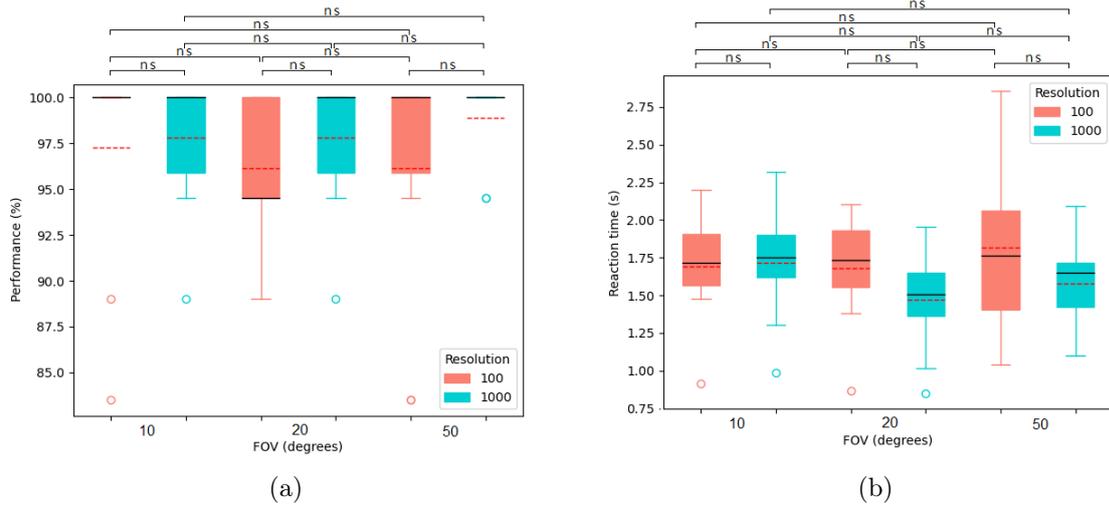


Figure 5.4: Light perception. Performance of correct responses and reaction time for Light module. (a) Box-plot for performance. (b) Box-plot for time.

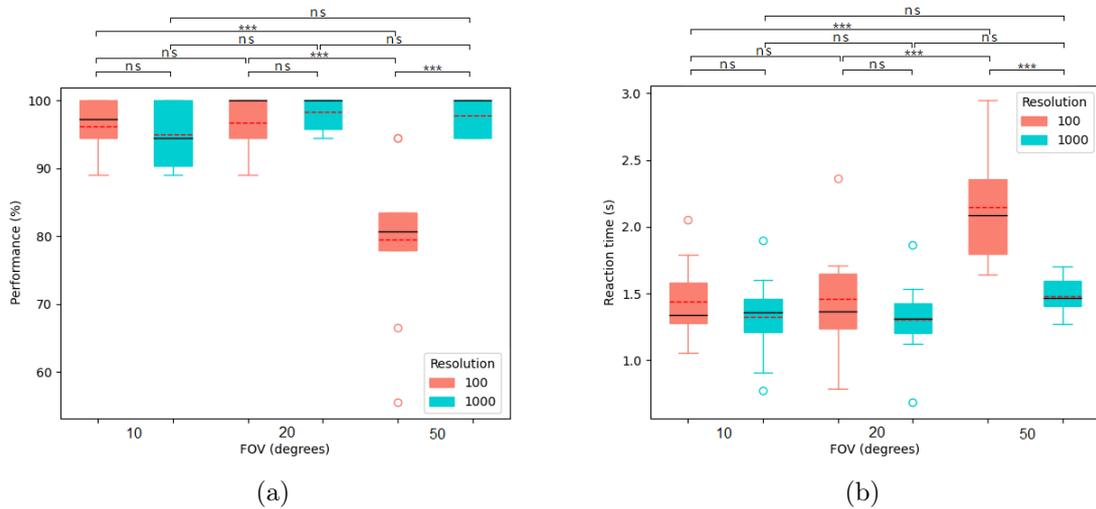


Figure 5.5: Time resolution. Performance of correct responses and reaction time for Time module. (a) Box-plot for performance. (b) Box-plot for time.

Figure 5.5(b) shows the reaction time for the ‘time resolution’ test. For the resolution of 100 phosphenes, the average reaction time is $1.44 \pm 0.31s$, $1.46 \pm 0.42s$ and $2.15 \pm 0.43s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.9947$). For the resolution of 1000, the average reaction time is $1.33 \pm 0.32s$, $1.31 \pm 0.30s$ and $1.48 \pm 0.14s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.9877$), 20-50 FOVs ($p=0.3160$) and 10-50 FOVs ($p=0.3909$).

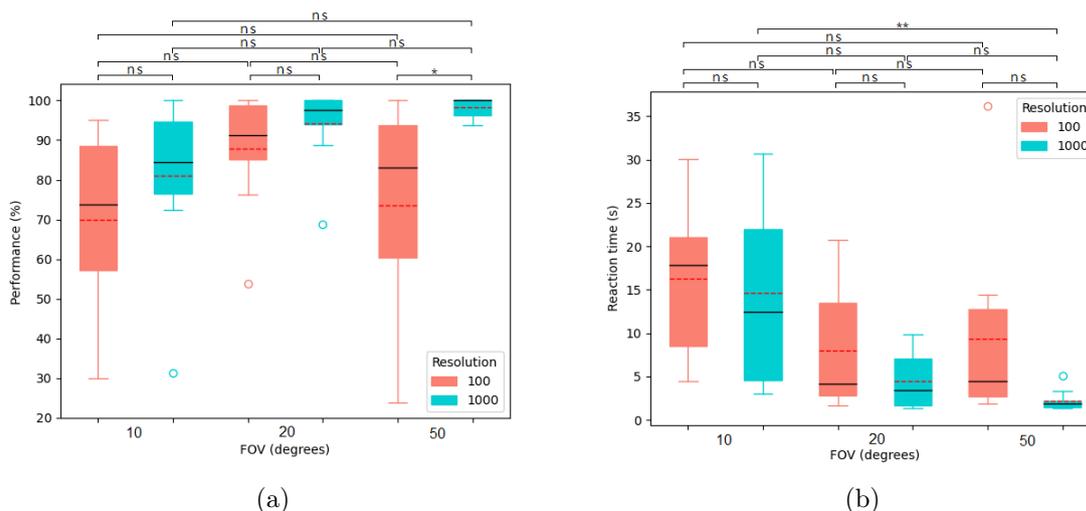


Figure 5.6: Light location. Performance of correct responses and reaction time for Location module. (a) Box-plot for performance. (b) Box-plot for time.

5.3.3 Light location

Figure 5.6 shows the performance and reaction time for the ‘light location’ test. The performance increases as the FOV and the resolution increases (see Figure 5.6(a)). For the resolution of 100 phosphenes the average performance is 70.00 ± 21.79 , 87.75 ± 14.37 and 73.63 ± 26.15 for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.1697$), 20-50 FOVs ($p=0.3158$) and 10-50 FOVs ($p=0.9237$). For the resolution of 1000 phosphenes the average performance is 81.00 ± 19.89 , 94.13 ± 9.70 and 98.38 ± 2.64 for 10, 20 and 50 degrees respectively. No significant difference was found for 20-50 FOVs ($p=0.7512$).

Figure 5.6(b) shows the reaction time for the three FOVs and two resolutions. For the resolution of 100 phosphenes the average reaction time is $16.25 \pm 8.63s$, $8.00 \pm 6.82s$ and $9.40 \pm 10.56s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.1099$), 20-50 FOVs ($p=0.9335$) and 10-50 FOVs ($p=0.2089$). For the resolution of 1000 phosphenes the average reaction time is $14.62 \pm 10.98s$, $4.47 \pm 3.27s$ and $2.24 \pm 1.17s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 20-50 FOVs ($p=0.6704$). Comparing the performance for the same FOV, the reaction time decreases with increasing number of phosphenes.

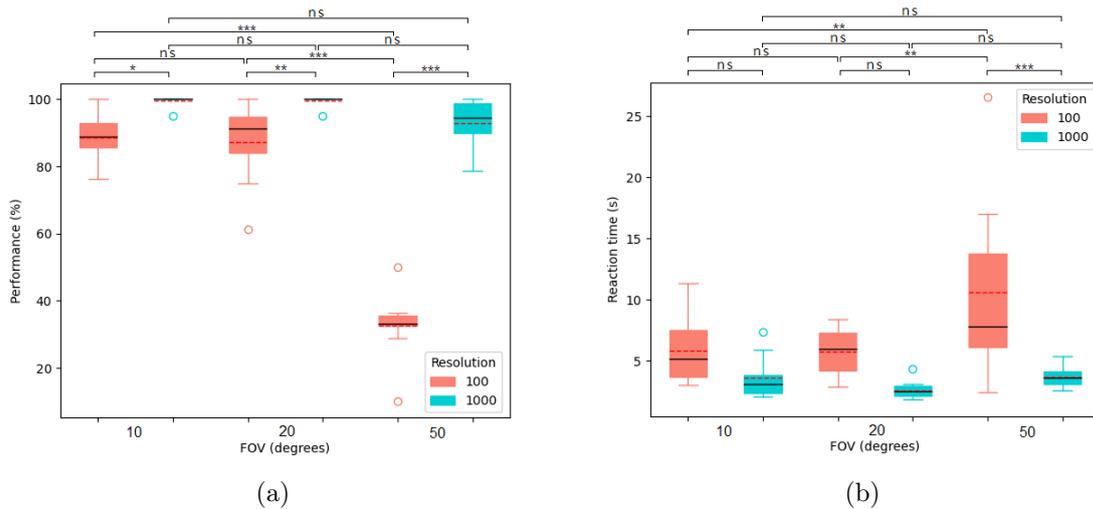


Figure 5.7: Motion perception. Performance of correct responses and reaction time for Motion module. (a) Box-plot for performance. (b) Box-plot for time.

5.3.4 Motion perception

Figure 5.7 shows the performance and reaction time for the ‘motion perception’ test. For the same resolution, the performance decreases as the FOV increases (see Figure 5.7(a)). For the resolution of 100 phosphenes the average performance is 88.63 ± 6.57 , 87.38 ± 11.64 and 32.63 ± 9.76 for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.9540$). For the resolution of 1000 phosphenes the average performance is 99.50 ± 1.58 , 99.50 ± 1.58 and 92.88 ± 7.17 for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=1.000$).

Figure 5.7(b) shows the reaction time for the ‘motion perception’ test. For the same resolution, the reaction time increases as the FOV increases. For the resolution of 100 phosphenes the average reaction time is $5.84 \pm 2.73s$, $5.80 \pm 1.94s$ and $10.59 \pm 7.13s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.9998$), 20-50 FOVs ($p=0.0647$) and 10-50 FOVs ($p=0.0675$). For the resolution of 1000 phosphenes the average reaction time is $3.64 \pm 1.72s$, $2.64 \pm 0.75s$ and $3.71 \pm 0.85s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.1591$), 20-50 FOVs ($p=0.1265$) and 10-50 FOVs ($p=0.9915$). Comparing the performance for the same FOV, the reaction time decreases with increasing number of phosphenes.

5.3.5 Landolt-C orientation

Figure 5.8 shows the performance and reaction time for the ‘Landolt-C orientation’ test. For the resolution of 100 phosphenes the average performance is 43.89 ± 23.78 , 59.44 ± 12.02 and 52.78 ± 11.19 for 10, 20 and 50 degrees respectively (see Figure 5.8(a)). No significant difference was found for 10-20 FOVs ($p=0.1141$), 20-50 FOVs ($p=0.6499$) and 10-50 FOVs ($p=0.4732$). For the resolution of 1000 phosphenes the average performance is 60.65 ± 10.94 , 63.33 ± 13.41 and 57.22 ± 7.43 for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.8366$), 20-50 FOVs ($p=0.4314$) and 10-50 FOVs ($p=0.7738$).

For the resolution of 100 phosphenes the average reaction time is $15.92 \pm 6.53s$, $8.92 \pm 3.93s$ and $9.24 \pm 2.70s$ for 10, 20 and 50 degrees, respectively. No significant difference was found for 20-50 FOVs ($p=0.9874$) (see Figure 5.8(b)). For the resolution of 1000 phosphenes the average reaction time is $8.29 \pm 2.91s$, $8.17 \pm 3.97s$ and $7.46 \pm 4.16s$ for 10, 20 and 50 degrees respectively. No significant difference was found for 10-20 FOVs ($p=0.9970$), 20-50 FOVs ($p=0.9041$) and 10-50 FOVs ($p=0.8711$).

Figure 5.9 shows the values of visual acuity (in logMAR) for the ‘Landolt-C orientation’ test obtained for each condition. Comparing the same FOV, we obtain higher visual acuity for 1000 resolution than for 100. However, the most significant difference between the two resolutions was found for the 10 FOV. The best visual acuity is obtained for the condition of 20° and 1000 resolution, with a visual acuity value of 1.3 logMAR. This visual acuity is considered the limit of blindness.

5.4 Discussion

During the past decade, much effort has been put into developing visual prosthesis devices that help restore vision in blind people. Nevertheless, less research has been done to find acceptable procedures for evaluating the functionality of different visual implant technologies and maximizing the benefits of machine vision.

Visual acuity tests have proven to be one of the main quantitative measures used to evaluate the efficacy and cost-effectiveness of procedures designed to improve or restore vision. Visual acuities of 0.5 logMAR and 1.0 logMAR are considered moderate and severe visual impairment, respectively. A visual acuity greater than 1.3 logMAR is considered

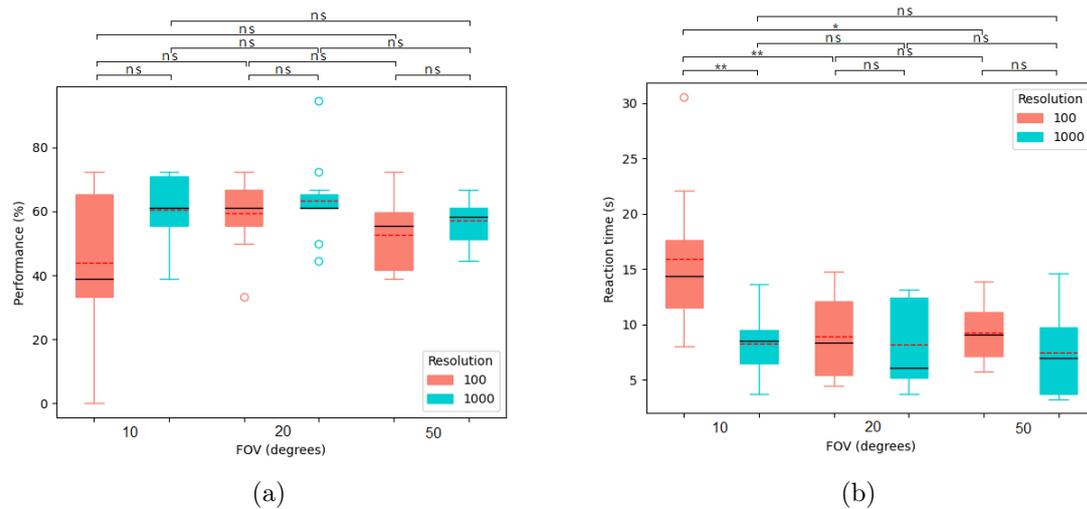


Figure 5.8: Landolt-C orientation. Performance of correct responses and reaction time for FrAct module. (a) Box-plot for performance. (b) Box-plot for time.

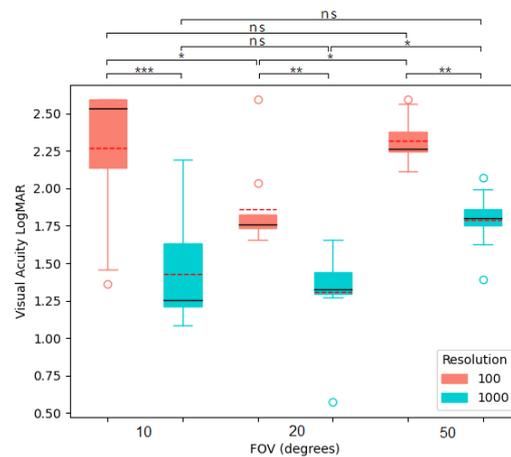


Figure 5.9: Visual acuity for Landolt-C. Values of visual acuity (in logMAR) for the ‘Landolt-C’ stimuli and for each condition.

total blindness. The theoretical visual acuity achievable by present-day retinal implants such as Argus II with a FOV of approximately 20° is 2.5 logMAR ($20/6325$), and the highest acuity is shown to be 1.8 logMAR ($20/1262$) [156, 161] in several basic tasks, such as those concerning grating visual acuity, square localization, and movement detection [81, 158–161]. Clinical trials for the Alpha-IMS assessed the visual acuity and object recognition of the restored vision in the context of daily living and mobility [30]. Three patients were able to read letters with a visual angle between 5° and 10° , demonstrating the highest visual acuity of 1.44 logMAR in the Landolt-C test. The visual acuity achieved by the Alpha-IMS is not significantly higher than that of the Argus II, given its 25-fold higher number of stimulating channels. These results may imply that the patterned

Table 5.1: Pixel density on the phosphenic image. Amount of pixels needed to form one phosphene on each of the conditions.

Resolution	Field of view (degrees)	Pixel density (Pixels/Phos)
100	10	9
	20	36
	50	361
1000	10	1
	20	4
	50	36

stimulation generated from all the 1500 individual channels of the Alpha-IMS could not be perfectly discriminated by the retinal cells, lowering the perceived spatial.

Visual acuity has already been used in the past both in clinical trials of visual neuroprosthesis [181, 277] but also in simulation experiments of prosthetic vision (SPV) [95, 185, 273]. The SPV software samples the image to match the resolution of the retinal prosthesis devices. By using SPV, researchers evaluated whether and under what conditions a measured visual acuity level is a true indication that the visual prosthesis provides a patterned image. Visual acuity is classically measured by optotypes such as letters, numbers or Landolt C-rings. Just as the interest in developing a visual prosthesis intensified in recent years, some researchers published their SPV study findings regarding the number of phosphenes required for comparable-to-normal visual acuity. In 2004, Chen et al. [273] examined visual acuity under virtual-reality (VR) in SPV using different filtering schemes. The best mean score recorded by the subjects was 1.55 logMAR. Later, they tested visual acuity for both rectangular and hexagonal phosphene grids using the Freiburg test [276]. The visual acuity scores ranged from 1.45 to 1.80 logMAR depending on subject. Similarly, Cha et al. [185] measured visual acuity as a function of the number of pixels and their spacing. They concluded that 625 electrodes implanted in a 1x1 cm area near the foveal representation of the visual cortex should produce a phosphene image with a visual acuity of approximately 20/30, 0.17 logMAR. Hayes et al. [95] simulate three retinal implants and test the functionality of this vision with four-choice orientation discrimination of a Sloan letter E. Subjects were found to have visual acuities of 1.96, 1.82, and 1.32 logMAR with the 4x4, 6x10, and 16x16 electrode arrays, respectively.

We found that the recognition of the different stimuli is well achieved with low resolution and restricted FOV. As can be seen in Figures 5.4, 5.5, 5.6, 5.7 and 5.8, for almost all

tests a significant improvement in task performance was obtained for a 20 FOV and a resolution of 1000. Besides, participants took less time to recognize the stimuli with this condition. Generally, participants took less time to recognize stimuli with the 20 FOV than the 50 FOV. This seems counterintuitive since with a narrower FOV the global reference of the image is lost. However, the narrower the FOV, the higher the angular resolution and therefore the greater the image detail (higher frequencies). Contrary, to generate one phosphene in the largest FOV, a larger number of pixels is averaged and therefore more information on image details is lost (see Table 5.1). Thus, the widest FOV allows to cover the widest area of the image but it only allows to see the gist of the image (low spatial frequency). On the other hand, the higher visual acuity was also obtained with the condition of 20 FOV and 1000 resolution (see Figure 5.9). For this condition, subjects obtained a visual acuity of 1.3 logMAR. If the number of phosphenes is reduced to 100 for the 20 FOV condition, visual acuity decreases to 1.86 logMAR. We can compare this visual acuity value with those obtained by some studies with Argus II using similar conditions of FOV and resolution [156, 161]. Furthermore, for the experimental condition of 10° FOV and 1000 resolution which can be compared to the Alpha-IMS implant [30], subjects obtained a visual acuity of 1.43 logMAR.

However, visual acuity is not the only important parameter nor the spatial details of visual scenes. There are many other relevant aspects in visual scenes such as shape, color and movement that would allow the extraction of complex information, for example identifying human faces, from relatively poor-quality images by using specific cues and multiple visual features [278] or obstacle detection from depth information and motion cues to facilitate the safe movement of the user in complex or unfamiliar environments [101]. This suggests that besides image resolution, we should try to pay attention to other relevant visual attributes such as receptive field size, localization, orientation, or movement [279]. In addition, depending on the subjects, there is one need or another. For example, some people focus more on identifying objects or people, while others prefer orientation and mobility. The key issue is to encode and send useful information that can be translated into functional gains for activities of daily living. Furthermore, it has been observed that there may be subtle differences in perceived visual field or encoding between subjects. Therefore, future advanced systems for interacting with the brain of people with low vision should allow the customization of functions to meet the needs of each subject.

5.5 Conclusions

Visual acuity tests are the main quantitative measures used to evaluate the effectiveness and cost-effectiveness of procedures designed to improve or restore vision. However, finding acceptable procedures for evaluating the functionality of visual implant technologies and maximizing the benefits of prosthetic vision is still under study. The present work constitutes a first essential step towards immersive virtual-reality simulations of prosthetic vision which has the potential to accelerate the prototyping of new devices. Via a head-mounted display, subjects were afforded simulated prosthetic vision (phosphene images) and required to recognise different stimuli normally used to measure visual acuity. Of all conditions tested, a FOV of 20° and 1000 phosphenes of resolution proved optimal, with higher visual acuity of 1.3 logMAR. Our simulated prosthetic vision system allows simple experimentation in order to study the design parameters of future visual prostheses. This work is a step toward the design of more effective electrode arrays that we hope will benefit the blind through neuroprosthesis.

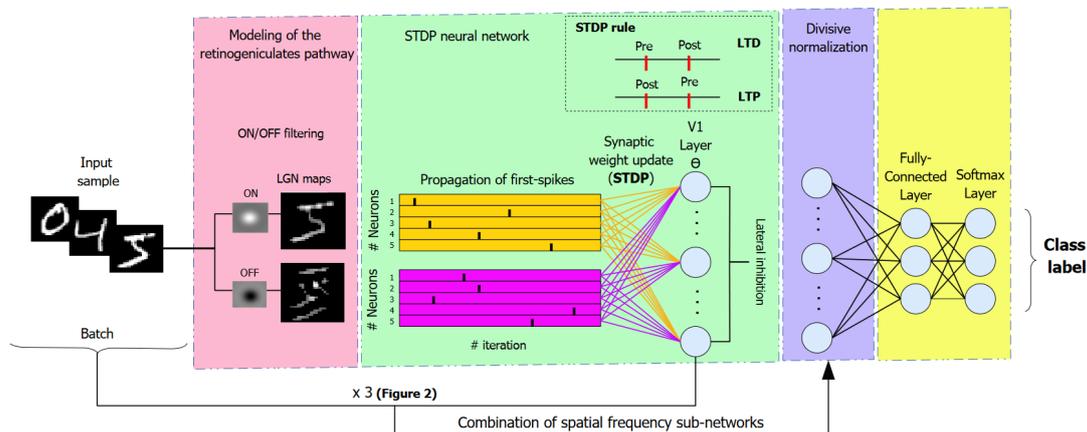
5.6 Related Publications

1. Sanchez-Garcia M., Morollón R., Martinez-Cantin R., and Guerrero J. J., Fernandez-Jover E. 2021. “Visual acuity assessment with visual prosthesis through a virtual-reality system”. Submitted.

Chapter 6

6 Spiking neural network

In this part of the thesis, we present a Spiking Neural Network (SNN) which relies on biologically plausible mechanisms and uses an unsupervised learning scheme to be part of a visual prostheses. The proposed SNN model can make use of the down-sampled signal from information processing unit of retinal prostheses bypassing retinal image analysis, providing useful information to the blind. Our SNN is separated in two stages. First, we simulated neurons in primary layers of the system which are capable of perceiving different spatial frequencies, similar to those found in the earliest stages of visual processing. These neurons fire asynchronously, with the most strongly activated neurons firing first. Second, we simulated neurons at later stages of the system, called V1 neurons, implementing a spike-timing-dependent-plasticity rule. We demonstrate that with one SNN layer V1 neurons become selective to features that are present in the input image. When presented with different spatial frequency images, V1 neurons detect coarse and fine information leading to a global representation of the scene. Our network also achieves an average image classification accuracy of 94.0% with the MNIST dataset and 82.0% for fashion-MNIST dataset.



6.1 Introduction

Current retinal prostheses have demonstrated significant progress in improving visual acuity and activities of daily living in blind patients. However, several technical problems are slowing down its evolution. One of the solutions being proposed is to develop an advanced retinal prosthesis to directly stimulate retinal ganglion cells (RGCs) with an array of electrodes using better computational models. For example, the reconstruction of visual scenes could be significantly improved by adding an encoder that converts the input images into the spiking codes used by RGCs; these codes are then used to drive electrodes.

The biological brain displays powerful information processing functionality. The primary visual cortex is the first stage in the processing of visual stimuli. There, neurons are synchronized in tune with a limited range of spatial frequencies (SF) ([280–284]). Neuron selective sensitivity for certain narrow ranges of SF is an immediate consequence of their receptive-field organization. Because neurons' receptive fields vary in size, the responses of different subsets of neurons would constitute a neural representation at some particular SF. As a result, neurons sensitive to low spatial frequencies are responsible for the global processing of visual scenes (the breaking-up of the scene into objects) and the neurons sensitive to high spatial frequencies are responsible for the local processing (fine details, recognition of particular patterns). However, our visual world is generally highly complex, composed of numerous features at a variety of scales, thereby having broad band SF spectra. This means that information signaled by individual neurons is highly incomplete, and combining information across multiple SF bands must be essential for the visual system to function in a robust and reliable manner.

In recent years various neural network models have been proposed to explain how the visual system is able to process an image efficiently ([285–288]). Specifically, Spiking Neural Networks (SNNs) have become an increasingly active field of research. This has been driven both by the interest to build more biologically realistic neural network models, and by recent improvements and the availability of larger-scale neuromorphic computing platforms, which emulate brain-like spike-based computation in dedicated analog or digital hardware ([285, 288–290]). On the other hand, SNNs are widely used because information can be transmitted using very weak signals, since rate coding is very resistant to noise,

and because they provide new learning algorithms for unsupervised learning. In fact, spike neurons allow the implementation of bioinspired local learning rules such as Hebbian learning and spike time-dependent plasticity (STDP) [291]. Also, it takes into account the times of the presynaptic and postsynaptic peaks and can be modified to achieve non-bimodal distributions of synaptic weights. The STDP rule is used in SNN to extract visual features of low or intermediate complexity in an unsupervised manner. From this, significant efforts have been expended in the recent past to demonstrate the efficacy of SNNs in pattern recognition applications. Yu et al. [292] proposed a novel spiking neural network with supervised learning rule and temporal coding scheme to generate the spike pattern. Such SNN system and its supervised learning rule achieved a relatively high correct classification rate when cross-validate the MNIST database. On the other hand, Liu et al. [293] a spike timing-based feed-forward spike neural network and its own unsupervised STDP learning rule achieving satisfactory results on MNIST database. Masquelier et al. [294] used the STDP rule in an asynchronous feedforward SNN that mimics the ventral visual pathway and showed that when the network was presented with natural images, selectivity to intermediate-complexity visual features emerged. Sboev et al. [295] proposed a method to solve the classification task using a SNN with encoding the input by patterns of spike times along with STDP learning. Furthermore, Diehl et al. [296] used STDP with lateral inhibition and rate coding of input data on the MNIST dataset, where digits were clustered into populations of neurons and then each cluster was labeled with corresponding digit.

We present a Spiking Neural Network (SNN) which relies on biologically plausible mechanisms and uses an unsupervised learning scheme to be part of a visual prosthesis. The proposed SNN model can make use of the down-sampled signal from the information processing unit of retinal prostheses bypassing retinal image analysis, providing useful information to the blind. Our SNN is separated in two stages. First, we simulated neurons in primary layers of the system which are capable of perceiving different spatial frequencies, similar to those found in the earliest stages of visual processing. These neurons fire asynchronously, with the most strongly activated neurons firing first. The first spikes approach is important for rapidly encoding the most important information about the visual stimulus. Second, we simulated neurons at later stages of the system, called V1 neurons, implementing a STDP rule which adjusts the efficacy of synaptic connections

based on the relative timing of post-synaptic spike and its input pre-synaptic spike.

6.2 Spiking Neural Network

Spiking neural networks are artificial neural networks that more closely mimic natural neural networks and are inspired by information processing in biology. They are being explored for their potential energy efficiency resulting from sparse, event-driven computations. In contrast to Artificial Neural Networks (ANNs) where information is presented by real numbers, information in SNNs is presented in the form of electrical impulses called spikes. The spikes are used as the only communication mechanism between network components which allow a complete desynchronization of the system because each component is only affected by the incoming spikes. The spike is described as a binary event, which can be defined by two parameters: the timestamp and the voltage.

The next sections describe the most important components for a SNN model.

6.2.1 Spiking Neurons

Spiking neurons are an intermediate model between biological neurons and artificial neurons. Various models of spiking neurons have been proposed ([297–300]). However, the most used spiking neuron model is the Leaky Integrate and Fire model (LIF) ([301]). This model integrates input spike to its membrane potential adding also a leak which allows neurons to return to the resting state in the absence of activity. LIF can be expressed as:

$$\tau_m \frac{dV}{dt} = -(V - E_L) + \frac{I}{g_L} \quad (6.1)$$

where V is the membrane potential, g_L is the leak conductance, E_L is the resting potential, I is the external input current, and τ_m is membrane time constant.

6.2.2 Neural coding

The neural coding in SNN is an important aspect to represent the spike mechanism. Typically, neural coding schemes are used to convert input pixels into spikes that are transmitted to the excitatory neurons. Temporal coding is the most used neural coding.

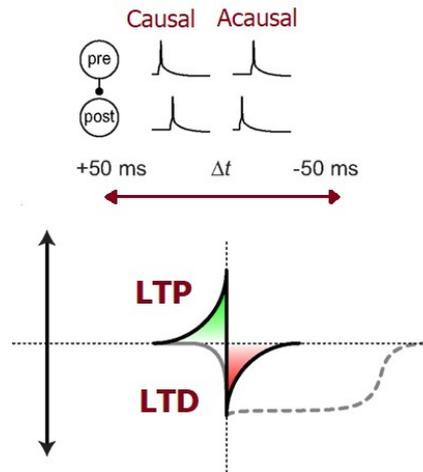


Figure 6.1: Spike-timing-dependent-plasticity (STDP) is the most studied learning rule for SNNs. STDP follows the Hebbian principle and suggests that synaptic connection depends on the difference of spike timestamps. (Adapted from Markram et al. [302]).

It represents the information on the timing of each spike. Thus, the value that encodes a spike is the offset of the spike timing according to a time reference, which locates the beginning of the pattern.

6.2.3 Synapses

Synapses can be present on the connection between two neurons. Their role is to modulate the voltage of the spikes transmitted on this connection, in order to modulate the influence of the input neuron over the output neuron. The modulation factor is defined by the synapse weight, a weak synaptic weight i.e. close to zero, will greatly reduce the influence of the input neuron over the output neuron, since spikes that reach the output neuron will have a low voltage (i.e. behave the same way as if there was no spike at all). On the contrary, a strong weight will produce post-synaptic spikes with high voltages, which will significantly affect the output neuron state. Adapting the weight of the synapse in the network directly affects the pattern that will excite the neurons, and thus the tasks that the network is able to solve. Training a network consists notably to make this adaptation, thanks to learning rules.

Spike-timing-dependent-plasticity (STDP) is the most studied learning rule for SNNs. STDP follows the Hebbian principle and suggest that synaptic connection depends on the difference of spike timestamps, following the next equation (6.2).

$$\Delta_w = \begin{cases} A_w e^{-\frac{t_{pre}-t_{post}}{\tau_{STDP}}}, & \text{if } t_{pre} \leq t_{post} \\ -A_w e^{-\frac{t_{post}-t_{pre}}{\tau_{STDP}}}, & \text{otherwise} \end{cases} \quad (6.2)$$

with A_w the learning rate and τ_{STDP} the time constant that controls the leak, t_{pre} and t_{post} , respectively the timestamp of fires for input and output neurons. This rule combines two mechanisms: the long-term potentiation (LTP), when the input neuron fires just before the output neuron, and the long-term depression (LTD) in the other case.

6.2.4 Inhibition

Inhibition means that spikes decrease the action potentials of the output neurons. That fact happens in biology, where excitatory neurons constitute about 80% and inhibitory neurons about 20%. A frequent case of use of inhibition is competition: when a neuron reacts to a pattern, it sends inhibitory spikes to other neurons in competition to prevent them from firing, and thus, increase the sparsity of the activity.

6.3 Method

In this section, the proposed feed-forward SNN model is presented and described in details along with the dataset and setup used in the experiments.

6.3.1 Dataset

In this work we use three datasets. We use a public database of indoor scenes from Quattoni et al. [140]. We choose five type of images, those belonging to indoor scene rooms: bathroom, bedroom, kitchen, living room and dining room. We test our network with this dataset for feature detection using different spatial frequencies. We also use the MNIST dataset from Lecun et al. [303] and the fashion-MNIST dataset from Xiao et al. [304]. Both, the MNIST and the fashion-MNIST datasets comprise 28x28 grayscale images of 70.000 digits or fashion products from 10 categories, with 7.000 images per category. The training set has 60.000 images and the test set has 10.000 images.

6.3.2 Pre-processing

Before input samples are converted into spikes to be fed to our SNN, we applied one pre-processing steps which is called “Modeling of the retinogeniculate pathway”. This first step corresponds to the primary processing of the visual information received by the eye’s retina. We focused on the lateral geniculate nucleus (LGN), where cells have center-surround receptive fields just like RGCs. We use a difference-of-Gaussians (DoG) filter to simulate on-off cells ([305]). The DoG filters detect contrasts in the input image in terms of spatial frequency (high, medium and low filter), as can be seen in Figure 6.2. To simulate the computations performed by the RGCs and the LGN, the images were convolved with ON and OFF center-surround kernels, followed by half-wave rectification. Thus, two maps were calculated, corresponding to ON- and OFF-center populations. The filters were implemented using difference of Gaussian (DoG) kernels, which were normalized to zero response for uniform inputs, and a response of 1 for the maximally stimulating input. These filters detect positive or negative contrasts in the input image and the result is an ON and OFF LGN activity maps. The receptive field sizes were the same as the size of the input image. The activity of each kernel can approximately be interpreted as a retinotopic contrast map of the sample. Further thresholding was applied such that, on average, only a small portion (the most active 10%) of the LGN units fired. This pre-processing allows the SNN to learn useful patterns. The LGN activity maps were converted to relative first-spike latencies through a monotonically decreasing function (a token function, $y = 1/x$, was chosen), but all monotonically decreasing functions are equivalent ([294]), thereby ensuring that the most active units (higher contrast) fired first, while units with lower activity (lower contrast) fired later or not at all. Latency-based encoding of stimulus properties has been reported extensively in the early visual system and it determines the content and the amount of information carried by each spike, which deeply affects the neural computations in the network ([306, 307]).

6.3.3 STDP neural network

We use a SNN with a STDP rule which extract visual features of low or intermediate complexity in an unsupervised manner. The samples were presented to the network sequentially. For each sample, the first spikes from the most active 10% of LGN neurons

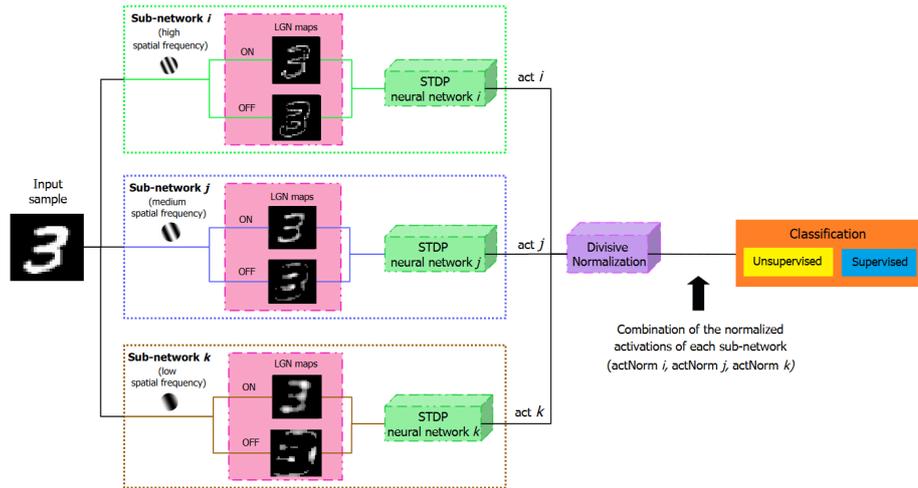


Figure 6.2: Combination of spatial frequency sub-networks. Our network combine the activation outputs of three sub-networks ($act\ i$, $act\ j$, $act\ k$), each corresponding to one spatial frequency: high, medium and low spatial frequency. Each of the activation vectors of the three sub-networks are normalized by a divisive normalization method. The normalized activations of the three networks ($actNorm\ i$, $actNorm\ j$, $actNorm\ k$) are combined into a single vector resulting in a vector of activations that combines different frequencies of activations. Finally, the classification is carried out using the unsupervised or supervised network.

were propagated through plastic synapses to a V1 population of 100 integrate-and-fire neurons. Furthermore, a winner-take-all inhibition scheme was implemented such that, if any V1 neuron fired during a certain iteration, it simultaneously prevented other neurons in the population from firing until the next sample was processed ([308]). After each iteration, the synaptic weights for the first V1 neuron to fire were updated using an STDP rule, and the membrane potentials of the entire population were reset to zero. This scheme leads to a sparse neural population where the probability of any two neurons learning the same feature is greatly reduced. At the start, each neuron was fully connected to all LGN afferents within its receptive field through synapses with randomly assigned weights between 0 and 1. The weights were restricted between 0 and 1 throughout the simulation. The non-negative values of the weights reflect the fact that thalamic connections to V1 are excitatory in nature ([309, 310]).

Each iteration began with the calculation of the ON and OFF LGN activity maps. This activity was thresholded, converted to spike latencies, and propagated through the network. The thresholding process allowed spikes from the fastest 10% of LGN neurons to propagate through the network. The propagated LGN spikes contributed to an increase in the membrane potential of V1 neurons (excitatory postsynaptic potentials) until one

of the V1 membrane potentials reached threshold, resulting in a postsynaptic spike and inhibition of all other V1 neurons until the next iteration (for more detail, see [291]).

After the LGN spike propagation, the synaptic weights were updated using an unsupervised multiplicative STDP rule [311]. The synaptic weight update was calculated as:

$$\Delta w = \begin{cases} -\alpha^- \cdot w^{\mu^-} \cdot K(\Delta t, \tau_-), & \Delta t \leq 0 \\ \alpha^+ \cdot (1 - w)^{\mu^+} \cdot K(\Delta t, \tau_+), & \Delta t > 0 \end{cases} \quad (6.3)$$

Here, a presynaptic and postsynaptic pair of spikes with a time difference (Δt) introduces a change in the weight (Δw) of the synapse, which is given by the product of a temporal filter K [typically, $K(\Delta t, \tau) = e^{-|\Delta t|/\tau}$] and a power function of its current weight (w). In our implementation, the efficiency and speed of calculations was greatly increased by making the windowing filter K infinitely wide (equivalent to assuming $\tau_{\pm} \rightarrow \infty$, or $K = 1$). This, however, does not imply that there was no temporal windowing in the model, as the thresholding at the LGN stage allowed only the fastest 10% of spikes to propagate in each iteration. As we said before, when a postsynaptic spike occurs shortly after a presynaptic spike ($\Delta t > 0$), there is a strengthening of the synapse, also called long-term potentiation (LTP). Conversely, when the postsynaptic spike occurs before the presynaptic spike ($\Delta t \leq 0$), or in the absence of a presynaptic spike, there is a weakening of the synapse or a long-term depression (LTD). The LTP and LTD are driven by their respective learning α^+ and α^- . The learning rates are non-negative ($\alpha^{\pm} \geq 0$) and determine the maximum amount of change in synaptic weights when $\Delta t \rightarrow \pm 0$. The parameters $\mu^{\pm} \in [0, 1]$ describe the degree of nonlinearity in the LTP and LTD update rules. In practice, a nonlinearity ensures that the final weights are graded and prevents convergence to bimodal distributions saturated at the upper and lower limits (0 and 1 in our case).

The propagated LGN spikes contributed to an increase in the membrane potential of V1 neurons (excitatory postsynaptic potentials or EPSPs) until one of the V1 membrane potentials reached threshold, resulting in a postsynaptic spike and inhibition of all other V1 neurons until the next iteration. The EPSP conducted by the synapse connecting the m th LGN neuron and the n th V1 neuron was taken as the weight of the synaptic connection itself (say w_{mn} wmn). The membrane potential, $E_n(t)$, of the n th V1 neuron

at time t is described as follow:

$$E_n(t) = \begin{cases} \sum_{m \in LGN} w_{mn} \cdot H(t - t_m), t < \min\{t \mid \max_{n \in V1} E_n(t) \geq \Theta\} \\ 0, otherwise \end{cases} \quad (6.4)$$

After a neuron fires the membrane potentials of the entire population were reset to zero. This scheme leads to a sparse neural population where the probability of any two neurons learning the same feature is greatly reduced. At the start, each neuron was fully connected to all LGN afferents within its receptive field through synapses with randomly assigned weights between 0 and 1. The weights were restricted between 0 and 1 throughout the simulation.

For the network used in the present study, the input images were presented sequentially. For each sample, the first spikes from the most active 10% of LGN neurons were propagated through plastic synapses to a V1 population of 100 integrate-and-fire neurons. At the start, each neuron was fully connected to all LGN afferents within its receptive field through synapses randomly assigned weights between 0 and 1. Here we used a winner-take-all inhibition scheme; if any V1 neurons fired during a certain iteration, it simultaneously prevent other neurons in the population from firing until the next sample. The learning rates were fixed with $\alpha^+ = 5 \times 10^{-3}$ and $\alpha^+/\alpha^- = (4/3)$. The rate ratio α^+/α^- is crucial to the stability of the network and was based on previous work demonstrating STDP-based visual feature learning ([294]). For these values of the learning rate, the threshold of the V1 neurons was fixed at $\Theta = 50$. This value was unmodified for all the results. Furthermore, we used a high nonlinearity for the LTP process ($\mu^+ = 0.65$) to ensure that we were able to capture fading receptive fields through continuous weights, and we used an almost adding LTD rule ($\mu^- = 0.05$) to ensure the pruning of continuously depressed synapses. In both LTP and LTD updates, the weights were maintained in the range $w \in [0,1]$. The STDP and inhibition rules were active only during the learning phase, and all subsequent testing of the converged population involved simultaneous spiking of all V1 neurons without any synaptic learning. After each iteration, the synaptic weights for the first V1 neuron to fire were updated using the Equation 6.3.

The post-convergence receptive fields of the STDP neurons were approximated by a linear summation of the afferent LGN receptive fields weighted by the converged synaptic weights. The receptive field ξ_j of the i th V1 neuron was estimated by the following:

$$\xi_i \approx \sum_{j \in LGN} w_{ij} \psi_j, \quad (6.5)$$

where ψ_j is the receptive field of the j th LGN afferent, and w_{ij} is the weight of the synapse connecting the j th afferent to the i th V1 neuron. The receptive fields calculated using this method are, in principle, similar to point wise estimates of the receptive field calculated by electrophysiologists.

6.3.4 Divisive normalization

The activation of the output neurons of each sub-network is normalized using a divisive normalization method. Let \bar{y}_{il} be the activation of neuron l in the sub-network i . The basic idea of divisive normalization is that the response of neuron l :

$$\bar{y}_{il}(x) = \frac{y_{il}^2(x)}{\sigma_i^2 + Wint_i \cdot \sum_{j \in K} y_j^2(x) + Wext_i \cdot \sum_{k \in K} y_k^2(x) + \sum_{i \in K} y_i^2(x)}, \quad (6.6)$$

is given by its driving input activity $y_{il}(x)$ in the sub-network i divisively normalized by a weighted sum over neurons' responses with various spatial frequency tunings ($y_j(x)$, $y_k(x)$) [312, 313], where x represents the stimulus and σ_i^2 is called the semi-saturation constant for sub-network i . As long as σ_i is nonzero, the normalized output will always be a value between 0 and 1, saturating for high contrasts. Here, the set of normalizing neurons K and the normalization weights, $Wint_i$ and $Wext_i$, define which neurons contribute to the normalization pool of neuron l and with what strength, respectively. With this model, the parameters σ_i , $Wint_i$ and $Wext_i$ are learned by optimizing the divisive normalization function using the gradient descent algorithm.

6.4 Results

In this section, we show the results of the classification task using an unsupervised and supervised networks.

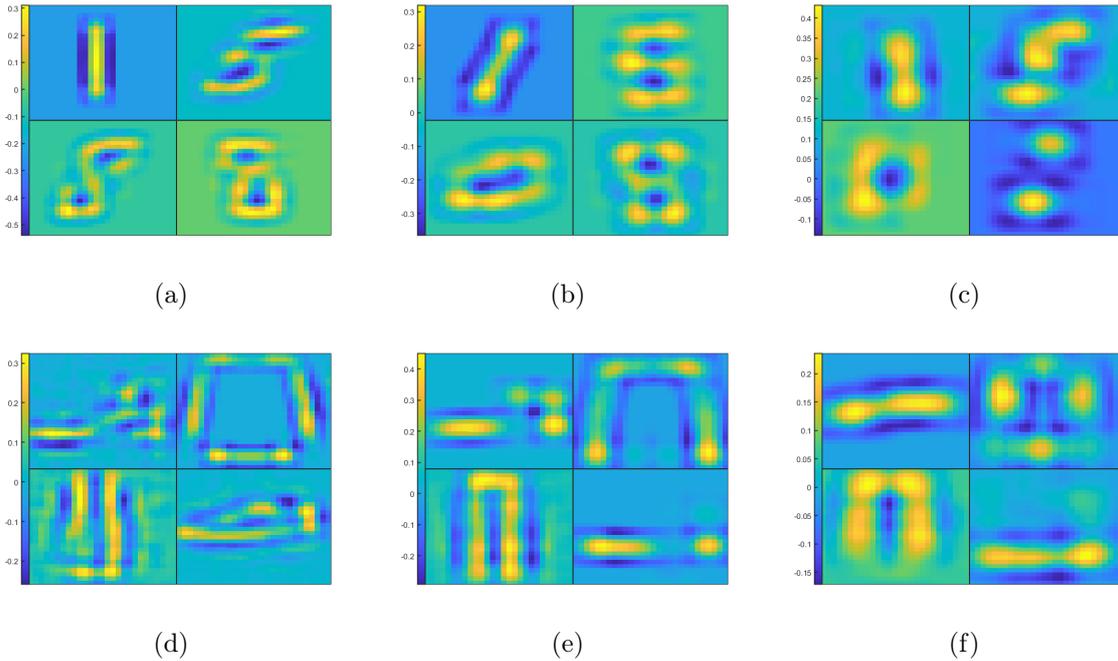


Figure 6.3: Receptive fields learned by different neurons for MNIST and fashion–MNIST [304] using various spatial frequencies (SFs). Top row: neurons receptive fields for MNIST, (a) high SF, (b) medium SF, (c) low SF. Bottom row: neurons receptive fields for fashion–MNIST, (d) high SF, (e) medium SF and (f) low SF.

6.4.1 Classification using supervised or unsupervised network

For both networks, the unsupervised and supervised, we train the network three times, each one corresponding to a specific spatial frequency or DoG filter: high pass, medium pass and low pass filter. We trained and tested each network with 100 excitatory neurons by presenting 40,000 examples of the MNIST and fashion–MNIST training set.

The learning rates were fixed with $\alpha^+ = 5 \times 10^{-3}$ and $\alpha^+/\alpha^- = (4/3)$. The rate ratio α^+/α^- is crucial to the stability of the network and was based on previous work demonstrating STDP-based visual feature learning [294]. For these values of the learning rate, the threshold of the V1 neurons was fixed at $\Theta = 50$. This value was unmodified for all the results. Furthermore, we used a high nonlinearity for the LTP process ($\mu^+ = 0.65$) to ensure that we were able to capture fading receptive fields through continuous weights, and we used an almost adding LTD rule ($\mu^- = 0.05$) to ensure the pruning of continuously depressed synapses. In both LTP and LTD updates, the weights were maintained in the range $w \in [0,1]$. The STDP and inhibition rules were active only during the learning phase, and all subsequent testing of the converged population involved simultaneous spiking of

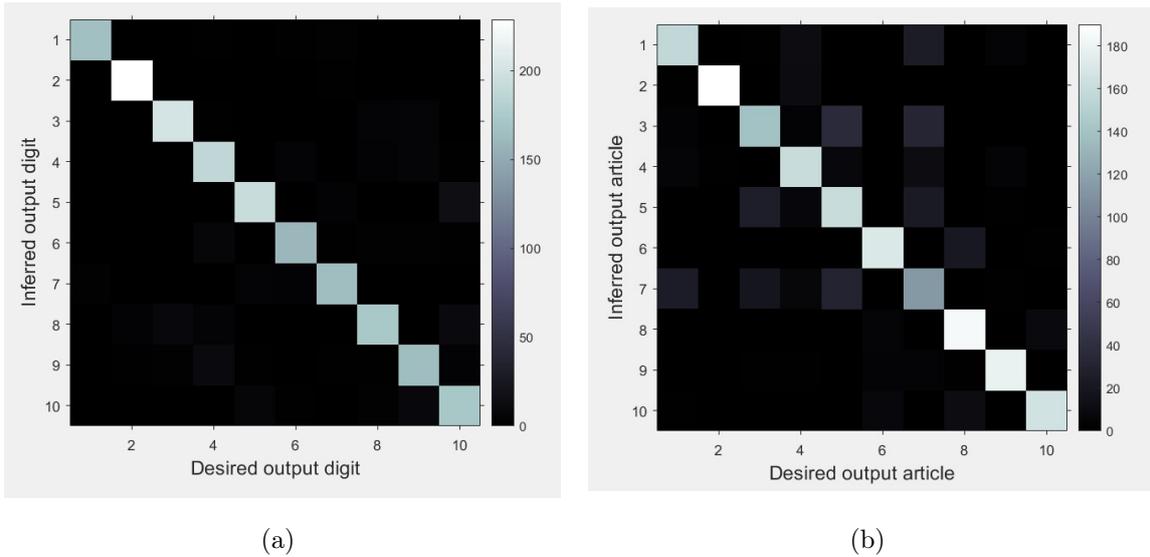


Figure 6.4: (a) Average confusion matrix of the testing results over ten presentations of the 10,000 MNIST test set digits. (b) Average confusion matrix of the testing results over ten presentations of the 10,000 fashion–MNIST test set articles.

all V1 neurons without any synaptic learning.

Figure 6.3 shows the preferred visual features of some neuronal maps in the first layer at the end of the learning process. The visual features learned by a neuron in the current layer is reconstructed as the weighted combinations of the visual features in the previous layer, whose preferred visual features are computed by DoG functions.

Before the classification layer, the normalized activations of the neurons of the three spatial frequencies sub-networks are combined into a single vector resulting in a vector of neuron activations with various spatial frequency tunings. To classify the images using the supervised approach, the normalized activation vector is fed to a fully-connected layer. In the output layer we use a softmax layer and compute the cross-entropy loss.

For the classification with the unsupervised network we use a classification method based on the average of spikes for each neuron. Once the training is done, we set the learning rate to zero, fix each neuron’s spiking threshold, and assign a class to each neuron, based on its highest response to the image class over the presentation of the training set. This is the only step where labels are used. The response of the class-assigned neurons is then used to measure the classification accuracy of the network on the test set. The predicted images are determined by averaging the responses of each neuron per class and then choosing the class with the highest average firing rate.

The two datasets, MNIST and fashion–MNIST achieved an average classification

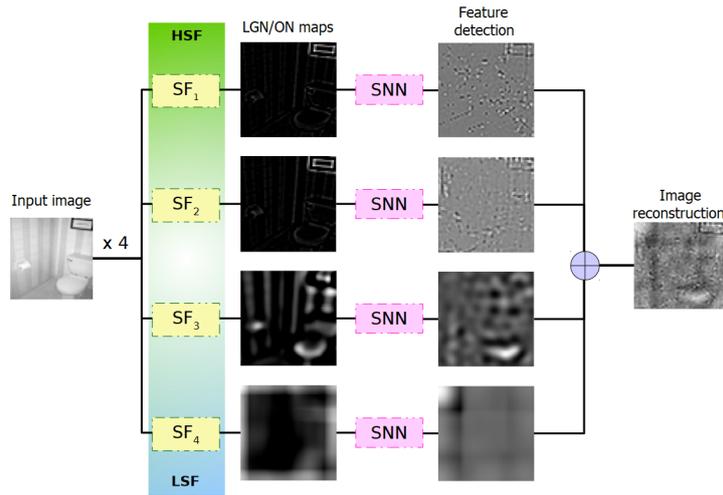


Figure 6.5: Feature detection using various spatial frequencies (SFs), from high (HSF) to low (LSF). Training our SNN individually using various SFs results in different types of features that, when combined, provide a sharper reconstruction of the image. Results using Quattoni et al. [140] dataset.

accuracy of 94.0 and 82.0% respectively. For both datasets we used the same neuron, synapse and STDP parameters.

Figure 6.4(a) shows the average confusion matrix over ten presentations of the MNIST test set, i.e., every single classification of the test examples belongs to one of the 10 by 10 tiles and its position is determined by the actual digit and the inferred digit. Not surprisingly, given a classification rate of 94%, most examples are on the identity which corresponds to correct classification; more interesting are the misclassified examples. The most common confusions are that 4 is 57 times identified as 9, 7 is identified 40 times as 9 and 7 is 26 times identified as 2. While 4 and 9, and 7 and 2 are easily confused it does not seem immediately obvious that a 7 could be mistaken as a 9. Often the only distinguishing feature between the misclassified 7's and a typical 9 is that the middle horizontal stroke in the 7 is not connected to upper stroke, which means that neurons which have a receptive field of a 9 are somewhat likely to fire as well. Figure 6.4(b) shows the average confusion matrix over ten presentations of the fashion–MNIST test set.

6.4.2 Feature detection

We also performed a second study for feature detection using different spatial frequencies with the unsupervised network. For the experiment we use the public database of Quattoni et al. [140]. We selected five types of indoor rooms: bathroom, bedroom, kitchen, dining

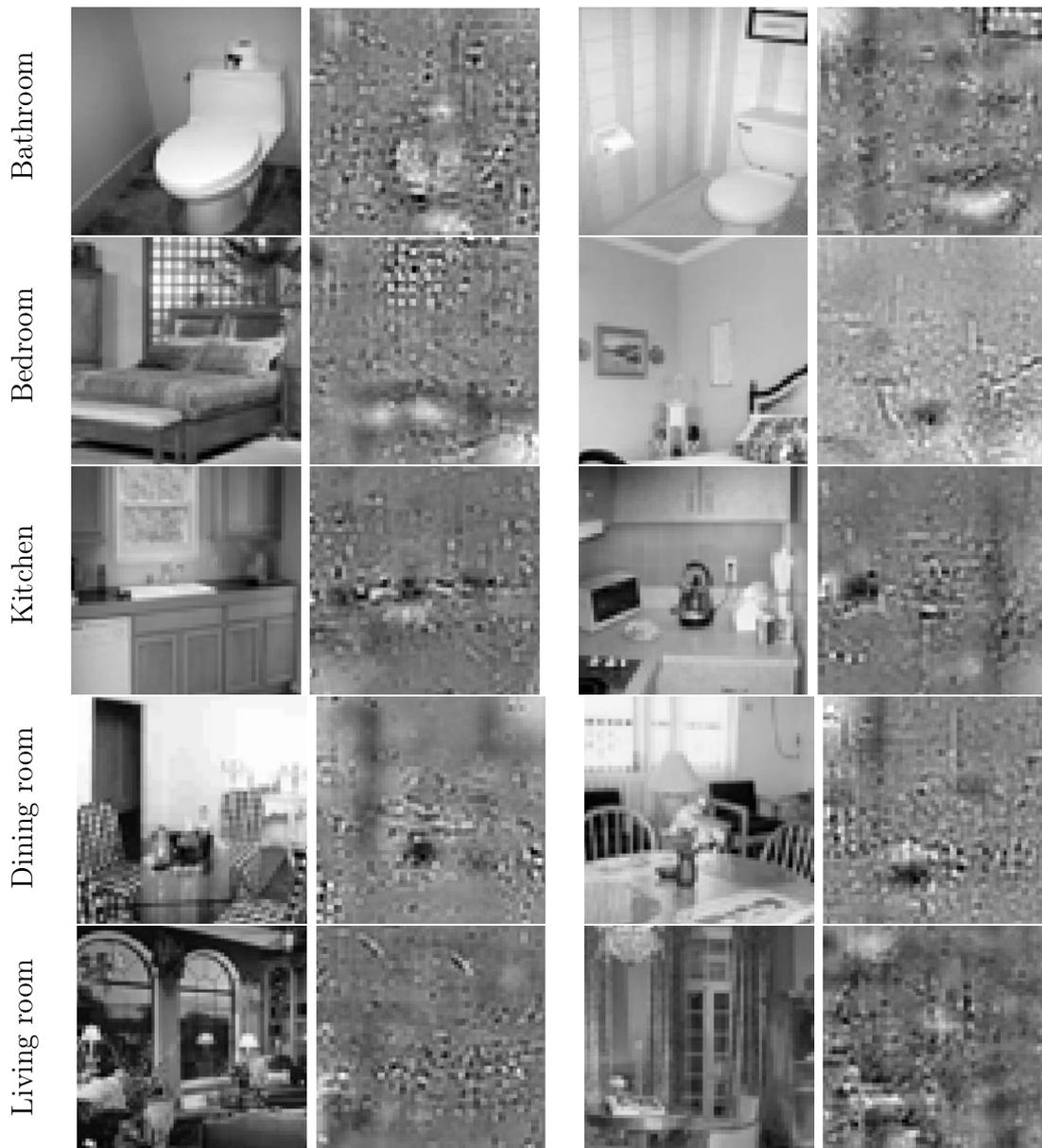


Figure 6.6: Results of features learned by our unsupervised network for Quattoni et al. [140] dataset. We show the features learned by our network in six examples of indoor scenes: bathroom, bedroom, kitchen, dining room and living room.

room and living room, as can be seen in Figure 6.6. We trained the unsupervised network with 250 images (50 images/class).

For the generation of the LGN maps, we processed the input image using four different spatial frequencies, from SF1 to SF4, being SF1 the highest spatial frequency and SF4 the lowest spatial frequency (see Figure 6.5). Figure 6.6 shows the image reconstruction based on the features detected by our network. For that, we use a linear approximation using the activation and receptive fields of each V1 neuron, as we can see in the equation:

$$IR_k = \sum_{j \in V1} a_{kj} \xi_j, \quad (6.7)$$

where a_{kj} is the activation value of j th V1 neuron in the k th image and ξ_j is the receptive field of the j th V1 neuron.

As we can see, training the network with different spatial frequencies, just as neurons do in primary layers of the system, the network detects different types of features that, when combined, provide a sharper reconstruction of the image (see Figure 6.6).

6.5 Conclusions

Theories on visual perception claims the existence of multiple channels, or multiple receptive field sizes, in the visual system the visual system. Because of the visual system has numerous relatively ‘narrow’ filters (channels) capable of perceiving different scales, scenes are processed in terms of spatial frequencies. This is important during scene recognition. Specifically, the primary visual system processes low and high level stimulus properties using inputs from the retina via the lateral geniculate nucleus. Low spatial frequencies, conveyed by fast magnocellular pathways, carry coarse information (e.g., the global shape and structure of a scene) whereas high spatial frequencies, conveyed more slowly by the parvocellular pathways, carry fine details of the scene (e.g., the edges and borders in the scene). We present a spiking neural network which relies on a combination of biologically plausible mechanisms and uses unsupervised learning scheme with multiple spatial frequencies. We demonstrate that with one spiking neural network layer V1 neurons become selective to features that are present in the input image. When presented with different spatial frequency images, V1 neurons detect coarse and fine information leading to a global representation of the scene. In the near future we want to test our network to classify more complex categories such as the Oliva et al. [314] dataset which is designed following principles of human visual cognition.

6.6 Related Publications

1. Sanchez-Garcia M., Chauchan T., Martinez-Cantin R., Guerrero J. J., and Cottureau B. 2022. “Spiking neural network using multiple spatial frequencies under an

unsupervised STDP Model”. In preparation.

Chapter 7

7 Conclusions and future prospects

To conclude this thesis, I would like to highlight some aspects for future research, which are particularly exciting for us, with the goal of improving future visual prostheses devices.

In this thesis we advance on various topics related to the design of visual prostheses but, of course, there is much more out there. In this thesis we have taken advantage of the fact that most retinal implants are equipped with an external video processing unit that is capable of applying simple image processing techniques to the visual input. In the last few years, deep learning has exploded, beating old traditional methods of the state of the art in many tasks. This allows the deployment of a diverse array of state-of-the-art deep-learning techniques for task-oriented and general image smart processing for the prosthesis, such as semantic segmentation and object detection, aimed at improving a patient's scene understanding. We have proposed to introduce deep learning, specifically fully convolutional networks to select and highlight useful information in interior scenes such as masks and silhouettes of relevant objects and structural edges that recover the main structure of the scene providing a sense of scale or perspective of the objects. Although deep learning methods are known to consume resources during training, they can achieve real-time performance for prediction even on mobile or embedded devices. Therefore, the next step in our research is to make our algorithm work in real time and easily integrate it into an implant device.

Although our results demonstrate the utility of deep learning-based scene simplification for prosthetic vision, there are a number of limitations that should be addressed in future work. First, we limited our approaches to indoor scenes. However, it would also be interesting to evaluate the performance of different scene simplification strategies in outdoor scenarios. For indoor scenes we use a structural and semantic segmentation algorithm for different layouts and types of objects, but these algorithms may perform differently compared to outdoor scenes. One of the problems that can occur in outdoor

environments is the interference of light and shadow contrasts.

Second, a major challenge is predicting what people ‘see’ when they use visual devices. Our studies simplify phosphenes into small independent light sources, although recent evidence suggests that phosphenes vary dramatically between subjects and electrodes. We can address this by incorporating a biologically realistic simulated prosthetic vision model in immersive virtual-reality, allowing sighted subjects to act as virtual patients in real-world tasks.

Third, the present study should be understood as a first step towards the ultimate goal of creating a retinal implant supported by deep learning–based image processing. Future devices would require all processing to happen in real time at the edge. Spiking Neural Networks (SNNs) with spike-timing-dependent-plasticity rule have been applied many times on image categorization tasks, in order to benchmark them against more common artificial neural networks. However, most of these studies used static images as stimuli. In contrast, dynamic images are more suited for SNNs due to their spatio-temporal nature. One solution could come in the form of low power, low-latency hardware such as an event-based vision camera. These cameras only encode variations in brightness or frequency and are totally asynchronous, just like the retina. That is why these sensors allow to take advantage of all the benefits of SNNs. Indeed, a combination of SNNs with event cameras could be a solution for tasks such as object detection, optical flow estimation or motion detection, among others. These studies show that a bioinspired system composed of an SNN driven by event-based camera inputs can learn, in an unsupervised manner, to optimally process spatio-temporal data. The ability to use accurate timing information is a feature only offered by SNNs, but has not been sufficiently explored and exploited. That is why SNNs have great potential in processing event-based sensor inputs, because only SNNs can fully exploit the precise temporal information that these sensors offer. To advance research in this direction, new benchmarks are needed, which do not carry the legacy of assessment in conventional machine learning or computer vision.

On the other hand, visual prosthetics are continually developing. At present, retinal prostheses are the most successful approach in this field. However, the inner layers of the retina can degenerate into many retinal diseases. Consequently, a retinal prosthesis may not be useful, for example if the inner layers of the retina are damaged, leading to advanced retinal degenerations such as glaucoma or optic atrophy. Thus, since the

neurons in the higher visual regions of the brain are usually spared from the damage to the retina and optic nerve, visual prostheses are being developed to directly stimulate the brain or designed for electrical stimulation of the visual cortex. Recently, Second Sight Medical Products has suspended production of new Argus II systems and has approved the Argus 2s Retinal Prosthesis System, a redesigned set of external hardware (glasses and video processing unit) initially for use in combination with previously implanted Argus II systems for the treatment of retinitis pigmentosa. The new Argus system will be adapted to be the external system for the next generation Orion Visual Cortical Prosthesis System currently under development. One of the benefits of shifting research to the new Orion Visual Cortical Prosthesis System is that, unlike the Argus II, which treats only retinitis pigmentosa, the Orion system treats a wide range of visual problems including glaucoma, diabetic retinopathy, optic nerve injury or disease, and eye injury.

To conclude, we could say that the development of visual prosthesis is still an emerging problem with many considerations to take into account, and many possible lines of research to move forward. This thesis proposes a collection of new methods that address some of the most important tasks related to scene understanding and object detection, but there are certainly others. The current climate of profound interest in computer vision and machine learning along with the progressive enhancements in technology to expect in the near future are very promising and encouraging to keep working on this important topic. We envision that the next steps in this direction of research will bring increased performance in the design of future visual prosthetics as well as a marked increase in the quality of life of people with visual impairments.

Conclusiones y perspectivas de futuro

Para concluir esta tesis, me gustaría destacar algunos aspectos para futuras investigaciones, que son especialmente emocionantes para nosotros, con el objetivo de mejorar los dispositivos de prótesis visuales del futuro.

En esta tesis avanzamos en diversos temas relacionados con el diseño de prótesis visuales pero, por supuesto, hay mucho más por hacer. En esta tesis hemos aprovechado que la mayoría de los implantes de retina están equipados con una unidad externa de procesamiento de video que es capaz de aplicar técnicas simples de procesamiento de imágenes a la entrada visual. En los últimos años, el aprendizaje profundo se ha disparado, superando los antiguos métodos tradicionales del estado del arte en muchas tareas. Esto permite el despliegue de una amplia gama de técnicas de aprendizaje profundo de vanguardia para el procesamiento inteligente de imágenes generales y orientado a tareas para la prótesis, como la segmentación semántica y la detección de objetos, con el objetivo de mejorar la comprensión de la escena del paciente. Hemos propuesto introducir el aprendizaje profundo, específicamente redes totalmente convolucionales para seleccionar y resaltar información útil en escenas de interior como máscaras y siluetas de objetos relevantes y bordes estructurales que recuperan la estructura principal de la escena proporcionando un sentido de escala o perspectiva de los objetos. Aunque se sabe que los métodos de aprendizaje profundo consumen recursos durante el entrenamiento, pueden lograr un rendimiento en tiempo real para la predicción incluso en dispositivos móviles o integrados. Por lo tanto, el siguiente paso en nuestra investigación es hacer que nuestro algoritmo funcione en tiempo real e integrarlo fácilmente en un dispositivo de implante.

Aunque nuestros resultados demuestran la utilidad de la simplificación de la escena basada en el aprendizaje profundo para la visión protésica, hay una serie de limitaciones que deben abordarse en el trabajo futuro. Primero, limitamos nuestros enfoques a escenas de interior. Sin embargo, también sería valioso evaluar el rendimiento de diferentes estrategias de simplificación de escenas en escenarios de exterior. Debido a que en las escenas de interior utilizamos un algoritmo de segmentación semántica y estructural para diferentes diseños y tipos de objetos, los algoritmos estudiados aquí pueden tener un

rendimiento diferente en comparación con las escenas de exterior. Uno de los problemas que pueden ocurrir en ambientes exteriores es la interferencia de contrastes de luces y sombras.

En segundo lugar, un desafío importante es predecir lo que las personas "ven" cuando utilizan dispositivos visuales. Nuestros estudios simplifican los fosfenos en pequeños puntos de luz independientes, aunque la evidencia reciente sugiere que los fosfenos varían drásticamente entre sujetos y electrodos. Podemos abordar esto incorporando un modelo de visión protésica simulada biológicamente realista en una realidad virtual inmersiva, lo que permite que los sujetos videntes actúen como pacientes virtuales en tareas del mundo real.

En tercer lugar, el presente estudio debe entenderse como un primer paso hacia el objetivo final de crear un implante de retina respaldado por un procesamiento de imágenes basado en el aprendizaje profundo. Los dispositivos futuros requerirían que todo el procesamiento se realice en tiempo real en el borde. Las redes neuronales de picos (SNN) con la regla de plasticidad dependiente del tiempo de picos se han aplicado muchas veces en tareas de categorización de imágenes, para compararlas con las redes neuronales artificiales más comunes. Sin embargo, la mayoría de estos estudios utilizaron imágenes estáticas como estímulos. Por el contrario, las imágenes dinámicas son más adecuadas para SNN debido a su naturaleza espacio-temporal. Una solución podría venir en forma de hardware de baja latencia y bajo consumo de energía, como una cámara de visión basada en eventos. Estas cámaras solo codifican variaciones de brillo o frecuencia y son totalmente asincrónicas, al igual que la retina. Es por eso que estos sensores permiten aprovechar todos los beneficios de las SNNs. De hecho, una combinación de SNN con cámaras de eventos podría ser una solución para tareas como detección de objetos, estimación de flujo óptico o detección de movimiento, entre otras. Estos estudios muestran que un sistema bioinspirado compuesto por un SNN impulsado por entradas de cámara basadas en eventos puede aprender, de manera no supervisada, a procesar de manera óptima datos espacio-temporales. La capacidad de utilizar información de sincronización precisa es una característica que solo ofrecen los SNN, pero no se ha explorado ni explotado lo suficiente. Es por eso que los SNN tienen un gran potencial en el procesamiento de entradas de sensores basados en eventos, porque solo los SNN pueden explotar completamente la información temporal precisa que ofrecen estos sensores. Para avanzar en la investigación

en esta dirección, se necesitan nuevos puntos de referencia, que no llevan el legado de la evaluación en el aprendizaje automático convencional o la visión por computadora.

Por otro lado, las prótesis visuales se están desarrollando continuamente. En la actualidad, las prótesis de retina son el enfoque más exitoso en este campo. Sin embargo, las capas internas de la retina pueden degenerar en muchas enfermedades de la retina. En consecuencia, una prótesis de retina puede no ser útil, por ejemplo, si las capas internas de la retina están dañadas, dando lugar a degeneraciones retinianas avanzadas como glaucoma o atrofia óptica. Por lo tanto, dado que las neuronas en las regiones visuales superiores del cerebro generalmente no sufren daños en la retina y el nervio óptico, se están desarrollando prótesis visuales para estimular directamente el cerebro o diseñadas para la estimulación eléctrica de la corteza visual. Recientemente, Second Sight Medical Products ha suspendido la producción de nuevos sistemas Argus II y ha aprobado el Sistema de Prótesis de Retina Argus 2s, un conjunto rediseñado de hardware externo (gafas y unidad de procesamiento de video) inicialmente para su uso en combinación con sistemas Argus II previamente implantados para el tratamiento de la retinitis pigmentosa. El nuevo sistema Argus está adaptado para ser el sistema externo del sistema de prótesis cortical visual Orion de próxima generación, actualmente en desarrollo. Uno de los beneficios de cambiar la investigación al nuevo sistema de prótesis cortical visual de Orion es que, a diferencia del Argus II, que solo trata la retinitis pigmentosa, el sistema Orion trata una amplia gama de problemas visuales que incluyen glaucoma, retinopatía diabética, lesiones o enfermedades del nervio óptico, y lesiones oculares.

Para concluir, podríamos decir que el desarrollo de prótesis visuales sigue siendo un problema emergente con muchas consideraciones a tener en cuenta, y muchas líneas de investigación posibles para avanzar. Esta tesis propone una colección de nuevos métodos que abordan algunas de las tareas más importantes relacionadas con la comprensión de la escena y la detección de objetos, pero ciertamente hay otras. El clima actual de profundo interés en la visión por computadora y el aprendizaje automático junto con las mejoras progresivas en la tecnología que se esperan en un futuro cercano son muy prometedores y alentadores para seguir trabajando en este importante tema. Prevemos que los próximos pasos en esta dirección de la investigación traerán un mayor rendimiento en el diseño de prótesis visuales futuras, así como un marcado aumento en la calidad de vida de las personas con discapacidad visual.

Chapter 8

8 Summary of results

8.1 Research Stays

During the years of the thesis I have made the following research stays:

- I was at the Institut de Recherche en Informatique de Toulouse (IRIT) and the Centre de recherche cerveau et cognition - CERCO from September 2019 to December 2019. I had the opportunity to meet and work with Prof. Benoit Cottureau and Dr. Tushar Chauchan.
- I also was at the Instituto de Bioingeniería, Universidad Miguel Hernandez Elche, from May 2021 to June 2021. My direct supervisor during this stay was Eduardo Fernández Jover.

8.2 Supervision of Students

Aside from research, I have also participated in teaching, imparting the “Visión y Robótica” and “Representación gráfica del patrimonio” subjects together with my supervisor Prof. JJ.Guerrero at the University of Zaragoza. Also, during the doctoral studies I participated in the supervision of students projects.

- Lorenzo Mur Labadia: “Bayesian deep learning for visual affordance”. MS. Robotics, CV, graphics 2020-2021
- Violeta Estepa Ramos: “Creación de un entorno visual de visión protésica”. MSc. Industrial Eng. 2019-2020
- María Santos Villafranca: “Simulador de prótesis visuales en entornos 360° con gafas de realidad virtual”. Bs. Industrial Eng. 2018-2019

- Pedro Luis Compais Serrano: “Sistema de asistencia a la navegación basado en la cámara de eventos”. MSc. Industrial Eng. 2018-2019

8.3 Dissemination

8.3.1 Peer-Reviewed Publications

The research developed in this thesis has resulted in the following peer-reviewed publications in journals, conferences and workshops.

Journals

- Sanchez-Garcia M., Martinez-Cantin R., Bermudez-Cameo J and Guerrero J. J. 2020. “Influence of field of view in visual prostheses design: Analysis with a VRsystem”. *Journal of Neural Engineering* (Q1)
- Sanchez-Garcia M, Martinez-Cantin R and Guerrero J. J, 2020. “Semantic and structural image segmentation for prosthetic vision”, *PLoS ONE* (Q2)
- Sanchez-Garcia M., Perez-Yus A., Martinez-Cantin R., and Guerrero J. J. 2021. “Augmented reality navigation system for visual prosthesis”. Submitted
- Sanchez-Garcia M., Morollon-Ruiz R., Martinez-Cantin R., Guerrero J. J., and Fernandez-Jover E. 2021. “Visual acuity assessment with visual prosthesis through a virtual-reality system.”. Submitted

In preparation

- Sanchez-Garcia M., Chauhan T., Martinez-Cantin R., Guerrero J. J., and Cottureau B. 2022. “Spiking neural network using multiple spatial frequencies under an unsupervised STDP Model”. In preparation.

Conferences

- Sanchez-Garcia M., Martinez-Cantin R., and Guerrero, J. J, 2019. “Indoor scenes understanding for visual prosthesis with fully convolutional networks”. In *14th International Conference on Computer Vision Theory and Applications*

Workshops

- Sanchez-Garcia M., Martinez-Cantin R., Bermudez-Cameo J and Guerrero, J. J, 2020. “Influence of Field of View in Visual Prostheses Design: Analysis with a VR System”. In *IX Jornada de Jovenes Investigadores (I3A)*
- Sanchez-Garcia M., Martinez-Cantin R., and Guerrero, J. J, 2019. “Semantic and Structural Image Segmentation for Prosthetic Vision”. In *VIII Jornada de Jovenes Investigadores (I3A)*
- Sanchez-Garcia M., Martinez-Cantin R., and Guerrero, J. J, 2018. “Smart Representation of Indoor Scenes under Simulated Prosthetic Vision”. In *WiCV Women in Computer Vision Workshop, 15th European Conference on Computer Vision (ECCV)*

8.3.2 Open-Source Software

We have released an Image dataset for SPV experimentation and also a virtual-reality software for replicating and extending SPV experimentation.

- **SIE-OMS dataset** (<https://doi.org/10.6084/m9.figshare.11493249.v4>)
- **SIE-OMS code** (<https://github.com/mesangar/SIE-OMS>)
- **Vrfov** (<http://webdiis.unizar.es/rmcantin/in-dex.php/Research/Vrfov>)

8.3.3 Conference and Research Seminar Attendance

I participated in the following courses or seminars offered outside the PhD program:

- Summer School on Computer Vision (BMVA 2018), Norwich - UK
- “Influence of Field of View in Visual Prostheses Design: Analysis with a VR System” Presented at IX Jornada de Jóvenes Investigadores del I3A, 2020
- I was featured in *Ágora*, which is the Aragón Radio program dedicated to popular science.
- “Semantic and Structural Image Segmentation for Prosthetic Vision” Presented at VIII Jornada de Jóvenes Investigadores del I3A, 2019

References

- [1] W. H. Organization, *Neurological disorders: public health challenges*. World Health Organization, 2006.
- [2] J.-W. Jeong, G. Shin, S. I. Park, K. J. Yu, L. Xu, and J. A. Rogers, “Soft materials in neuroengineering for hard problems in neuroscience,” *Neuron*, vol. 86, no. 1, pp. 175–186, 2015.
- [3] J. Rivnay, H. Wang, L. Fenno, K. Deisseroth, and G. G. Malliaras, “Next-generation probes, particles, and proteins for neural interfacing,” *Science Advances*, vol. 3, no. 6, p. e1601649, 2017.
- [4] M. D. Ferro and N. A. Melosh, “Electronic and ionic materials for neurointerfaces,” *Advanced Functional Materials*, vol. 28, no. 12, p. 1704335, 2018.
- [5] S. Luan, I. Williams, K. Nikolic, and T. G. Constandinou, “Neuromodulation: present and emerging methods,” *Frontiers in neuroengineering*, vol. 7, p. 27, 2014.
- [6] S. B. Brummer and M. Turner, “Electrochemical considerations for safe electrical stimulation of the nervous system with platinum electrodes,” *IEEE Transactions on Biomedical Engineering*, no. 1, pp. 59–63, 1977.
- [7] M. H. Maghami, A. M. Sodagar, A. Lashay, H. Riazi-Esfahani, and M. Riazi-Esfahani, “Visual prostheses: the enabling technology to give sight to the blind,” *Journal of ophthalmic & vision research*, vol. 9, no. 4, p. 494, 2014.
- [8] S. Brummer and M. Turner, “Electrical stimulation of the nervous system: the principle of safe charge injection with noble metal electrodes,” *Bioelectrochemistry and Bioenergetics*, vol. 2, no. 1, pp. 13–25, 1975.
- [9] T. Stieglitz, “Neuro-technical interfaces to the central nervous system,” *Poiesis & Praxis*, vol. 4, no. 2, pp. 95–109, 2006.
- [10] W. Tong, H. Meffin, D. J. Garrett, and M. R. Ibbotson, “Stimulation strategies for improving the resolution of retinal prostheses,” *Frontiers in neuroscience*, vol. 14, p. 262, 2020.
- [11] I. Rehman, B. Hazhirkarzar, and B. C. Patel, “Anatomy, head and neck, eye,” *StatPearls [Internet]*, 2019.
- [12] G. WHO *et al.*, “World health organization,” *Guidelines on Optimal Feeding of Low Birthweight Infants in Low-and Middle-Income Countries*, 2011.
- [13] J. C. Ten Berge, Z. Fazil, L. I. van den Born, R. C. Wolfs, M. W. Schreurs, W. A. Dik, and A. Rothova, “Intraocular cytokine profile and autoimmune reactions in retinitis pigmentosa, age-related macular degeneration, glaucoma and cataract,” *Acta ophthalmologica*, vol. 97, no. 2, pp. 185–192, 2019.
- [14] L. Wu, X. Sun, X. Zhou, and C. Weng, “Causes and 3-year-incidence of blindness in jing-an district, shanghai, china 2001-2009,” *BMC ophthalmology*, vol. 11, no. 1, pp. 1–6, 2011.

- [15] M. Singh and S. C. Tyagi, “Genes and genetics in eye diseases: a genomic medicine approach for investigating hereditary and inflammatory ocular disorders,” *International journal of ophthalmology*, vol. 11, no. 1, p. 117, 2018.
- [16] P. Yu-Wai-Man, N. J. Newman, V. Carelli, M. L. Moster, V. Biousse, A. A. Sadun, T. Klopstock, C. Vignal-Clermont, R. C. Sergott, G. Rudolph *et al.*, “Bilateral visual improvement with unilateral gene therapy injection for leber hereditary optic neuropathy,” *Science translational medicine*, vol. 12, no. 573, 2020.
- [17] K. Narfström, M. Katz, M. Ford, T. Redmond, E. Rakoczy, and R. Bragadottir, “In vivo gene therapy in young and adult rpe65^{-/-} dogs produces long-term visual improvement,” *Journal of Heredity*, vol. 94, no. 1, pp. 31–37, 2003.
- [18] M. Simunovic, W. Shen, J. Lin, D. Protti, L. Lisowski, and M. Gillies, “Optogenetic approaches to vision restoration,” *Experimental eye research*, vol. 178, pp. 15–26, 2019.
- [19] M. E. McClements, F. Staurenghi, R. E. MacLaren, and J. Cehajic-Kapetanovic, “Optogenetic gene therapy for the degenerate retina: recent advances,” *Frontiers in Neuroscience*, vol. 14, p. 1187, 2020.
- [20] R. K. Shepherd, M. N. Shivdasani, D. A. Nayagam, C. E. Williams, and P. J. Blamey, “Visual prostheses for the blind,” *Trends in biotechnology*, vol. 31, no. 10, pp. 562–571, 2013.
- [21] J. D. Weiland and M. S. Humayun, “Visual prosthesis,” *Proceedings of the IEEE*, vol. 96, no. 7, pp. 1076–1084, 2008.
- [22] B. Franklin, *Experiments and observations on electricity, made at Philadelphia in America... To which are added, letters and papers on philosophical subjects. The whole corrected, methodized... and now first collected into one volume, etc.* [Edited by Peter Collinson.]. David Henry, 1769.
- [23] C. LeRoy, “Où l’on rend compte de quelques tentatives que l’on a faites pour guérir plusieurs maladies par l’électricité,” *Hist Acad Roy Sciences Memoires Math Phys*, vol. 60, pp. 87–95, 1755.
- [24] N. R. Stiles, B. P. McIntosh, P. J. Nasiatka, M. C. Hauer, J. D. Weiland, M. S. Humayun, and A. R. TANGUAY, JR, “An intraocular camera for retinal prostheses: Restoring sight to the blind,” in *Optical Processes in Microparticles and Nanostructures: A Festschrift Dedicated to Richard Kounai Chang on His Retirement from Yale University*. World Scientific, 2011, pp. 385–429.
- [25] Y. H.-L. Luo and L. Da Cruz, “The argus® ii retinal prosthesis system,” *Progress in retinal and eye research*, vol. 50, pp. 89–107, 2016.
- [26] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [27] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, “Better speech recognition with cochlear implants,” *Nature*, vol. 352, no. 6332, pp. 236–238, 1991.

- [28] C. Feng, S. Dai, Y. Zhao, and S. Liu, "Edge-preserving image decomposition based on saliency map," in *2014 7th International Congress on Image and Signal Processing*. IEEE, 2014, pp. 159–163.
- [29] J. Loudin, A. Butterwick, P. Huie, and D. Palanker, "Delivery of information and power to the implant, integration of the electrode array with the retina, and safety of chronic stimulation," *Visual Prosthetics*, pp. 137–158, 2011.
- [30] K. Stingl, K. U. Bartz-Schmidt, D. Besch, C. K. Chee, C. L. Cottrill, F. Gekeler, M. Groppe, T. L. Jackson, R. E. MacLaren, A. Koitschev *et al.*, "Subretinal visual implant alpha-ims—clinical trial interim report," *Vision research*, vol. 111, pp. 149–160, 2015.
- [31] K. Stingl, K. U. Bartz-Schmidt, D. Besch, A. Braun, A. Bruckmann, F. Gekeler, U. Greppmaier, S. Hipp, G. Hörtdörfer, C. Kernstock *et al.*, "Artificial vision with wirelessly powered subretinal electronic implant alpha-ims," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1757, p. 20130077, 2013.
- [32] E. Bloch, Y. Luo, and L. da Cruz, "Advances in retinal prosthesis systems," *Therapeutic advances in ophthalmology*, vol. 11, p. 2515841418817501, 2019.
- [33] K. Stingl, R. Schippert, K. U. Bartz-Schmidt, D. Besch, C. L. Cottrill, T. L. Edwards, F. Gekeler, U. Greppmaier, K. Kiel, A. Koitschev *et al.*, "Interim results of a multicenter trial with the new electronic subretinal implant alpha-ams in 15 patients blind from inherited retinal degenerations," *Frontiers in neuroscience*, vol. 11, p. 445, 2017.
- [34] R. Hornig, M. Dapper, E. Le Joliff, R. Hill, K. Ishaque, C. Posch, R. Benosman, Y. LeMer, J.-A. Sahel, and S. Picaud, "Pixium vision: first clinical results and innovative developments," in *Artificial Vision*. Springer, 2017, pp. 99–113.
- [35] R. Daschner, A. Rothermel, R. Rudolf, S. Rudolf, and A. Stett, "Functionality and performance of the subretinal implant chip alpha-ams," *Sens. Mater*, vol. 30, no. 2, pp. 179–192, 2018.
- [36] A. P. Finn, D. S. Grewal, and L. Vajzovic, "Argus ii retinal prosthesis system: a review of patient selection criteria, surgical considerations, and post-operative outcomes," *Clinical Ophthalmology (Auckland, NZ)*, vol. 12, p. 1089, 2018.
- [37] M. Vurro, A. M. Crowell, and J. S. Pezaris, "Simulation of thalamic prosthetic vision: reading accuracy, speed, and acuity in sighted humans," *Frontiers in human neuroscience*, vol. 8, p. 816, 2014.
- [38] G. Dagnelie, P. Keane, V. Narla, L. Yang, J. Weiland, and M. Humayun, "Real and virtual mobility performance in simulated prosthetic vision," *Journal of Neural Engineering*, vol. 4, no. 1, p. S92, 2007.
- [39] L. Wang, L. Yang, and G. Dagnelie, "Virtual wayfinding using simulated prosthetic vision in gaze-locked viewing," *Optometry and vision science: official publication of the American Academy of Optometry*, vol. 85, no. 11, p. E1057, 2008.
- [40] J. Kasowski, N. Wu, and M. Beyeler, "Towards immersive virtual reality simulations of bionic vision," *arXiv preprint arXiv:2102.10678*, 2021.

- [41] A. Caspi, J. D. Dorn, K. H. McClure, M. S. Humayun, R. J. Greenberg, and M. J. McMahon, “Feasibility study of a retinal prosthesis: spatial vision with a 16-electrode implant,” *Archives of Ophthalmology*, vol. 127, no. 4, pp. 398–401, 2009.
- [42] A. Horsager, G. M. Boynton, R. J. Greenberg, and I. Fine, “Temporal interactions during paired-electrode stimulation in two retinal prosthesis subjects,” *Investigative ophthalmology & visual science*, vol. 52, no. 1, pp. 549–557, 2011.
- [43] M. Beyeler, D. Nanduri, J. D. Weiland, A. Rokem, G. M. Boynton, and I. Fine, “A model of ganglion axon pathways accounts for percepts elicited by retinal implants,” *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.
- [44] M. Beyeler, G. M. Boynton, I. Fine, and A. Rokem, “pulse2percept: A python-based simulation framework for bionic vision,” *BioRxiv*, p. 148015, 2017.
- [45] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell, “Simulating prosthetic vision: I. visual models of phosphenes,” *Vision Research*, vol. 49, no. 12, pp. 1493–1506, 2009.
- [46] D. Rushton and G. Brindley, “Properties of cortical electrical phosphenes,” in *Frontiers in visual science*. Springer, 1978, pp. 574–593.
- [47] J. Wang, Y. Lu, L. Gu, C. Zhou, and X. Chai, “Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies,” *Information Sciences*, vol. 277, pp. 512–524, 2014.
- [48] E. McKone, R. A. Robbins, X. He, and N. Barnes, “Caricaturing faces to improve identity recognition in low vision simulations: How effective is current-generation automatic assignment of landmark points?” *PloS one*, vol. 13, no. 10, p. e0204361, 2018.
- [49] M. J.-M. Macé, V. Guivarch, G. Denis, and C. Jouffrais, “Simulated prosthetic vision: the benefits of computer-based object recognition and localization,” *Artificial Organs*, vol. 39, no. 7, pp. E102–E113, 2015.
- [50] M. P. Zapf, P. B. Matteucci, N. H. Lovell, S. Zheng, and G. J. Suaning, “Towards photorealistic and immersive virtual-reality environments for simulated prosthetic vision: Integrating recent breakthroughs in consumer hardware and software,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 2597–2600.
- [51] F. I. Kiral-Kornek, E. OSullivan-Greene, C. O. Savage, C. McCarthy, D. B. Grayden, and A. N. Burkitt, “Improved visual performance in letter perception through edge orientation encoding in a retinal prosthesis simulation,” *Journal of neural engineering*, vol. 11, no. 6, p. 066002, 2014.
- [52] Y. Zhao, Y. Lu, C. Zhou, Y. Chen, Q. Ren, and X. Chai, “Chinese character recognition using simulated phosphene maps,” *Investigative ophthalmology & visual science*, vol. 52, no. 6, pp. 3404–3412, 2011.
- [53] X. Chai, W. Yu, J. Wang, Y. Zhao, C. Cai, and Q. Ren, “Recognition of pixelized chinese characters using simulated prosthetic vision,” *Artificial organs*, vol. 31, no. 3, pp. 175–182, 2007.

- [54] J. Sommerhalder, E. Oueghlani, M. Bagnoud, U. Leonards, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision: I. eccentric reading of isolated words, and perceptual learning," *Vision Research*, vol. 43, no. 3, pp. 269–283, 2003.
- [55] G. Dagnelie, D. Barnett, M. S. Humayun, and R. W. Thompson, "Paragraph text reading using a pixelized prosthetic vision simulator: parameter dependence and task learning in free-viewing conditions," *Investigative ophthalmology & visual science*, vol. 47, no. 3, pp. 1241–1250, 2006.
- [56] J. Sommerhalder, B. Rappaz, R. de Haller, A. P. Fornos, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision: II. eccentric reading of full-page text and the learning of this task," *Vision research*, vol. 44, no. 14, pp. 1693–1706, 2004.
- [57] L. Fu, S. Cai, H. Zhang, G. Hu, and X. Zhang, "Psychophysics of reading with a limited number of pixels: towards the rehabilitation of reading ability with visual prosthesis," *Vision Research*, vol. 46, no. 8-9, pp. 1292–1301, 2006.
- [58] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.
- [59] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.
- [60] G. Dagnelie, R. Thompson, G. Barnett, and W. Zhang, "Visual perception and performance under conditions simulating prosthetic vision," *Perception ECVF abstract*, vol. 29, pp. 0–0, 2000.
- [61] G. Dagnelie, R. W. Thompson, G. D. Baraett, and W. Zhang, "Simulated prosthetic vision: perceptual and performance measures," in *Vision Science and its Applications*. Optical Society of America, 2001, p. FC2.
- [62] M. Chang, H. Kim, J. Shin, and K. Park, "Facial identification in very low-resolution images simulating prosthetic vision," *Journal of Neural Engineering*, vol. 9, no. 4, p. 046012, 2012.
- [63] R. W. Thompson, G. D. Barnett, M. S. Humayun, and G. Dagnelie, "Facial recognition using simulated prosthetic pixelized vision," *Investigative ophthalmology & visual science*, vol. 44, no. 11, pp. 5035–5042, 2003.
- [64] J. Wang, X. Wu, Y. Lu, H. Wu, H. Kan, and X. Chai, "Face recognition in simulated prosthetic vision: face detection-based image processing strategies," *Journal of Neural Engineering*, vol. 11, no. 4, p. 046009, 2014.
- [65] G. Denis, M. Macé, and C. Jouffrais, "Simulated prosthetic vision: object recognition and localization approach," in *Proceedings of the 4th International Conference on Neuroprosthetic Devices (ICNPD 2012)*, 2012, pp. 40–1.
- [66] T. Han, H. Li, Q. Lyu, Y. Zeng, and X. Chai, "Object recognition based on a foreground extraction method under simulated prosthetic vision," in *Bioelectronics and Bioinformatics (ISBB), 2015 International Symposium on*. IEEE, 2015, pp. 172–175.

- [67] H. Li, X. Su, J. Wang, H. Kan, T. Han, Y. Zeng, and X. Chai, “Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision,” *Artificial Intelligence in Medicine*, vol. 84, pp. 64–78, 2018.
- [68] N. Han, S. Srivastava, A. Xu, D. Klein, and M. Beyeler, “Deep learning–based scene simplification for bionic vision,” *arXiv preprint arXiv:2102.00297*, 2021.
- [69] J. J. van Rheede, C. Kennard, and S. L. Hicks, “Simulating prosthetic vision: Optimizing the information content of a limited visual display,” *Journal of Vision*, vol. 10, no. 14, pp. 32–32, 2010.
- [70] TATUR-G., *Conception d un systeme de vision par phosphenes*. UNIV EUROPEENNE, 2011.
- [71] P. Lieby, N. Barnes, C. McCarthy, N. Liu, H. Dennett, J. G. Walker, V. Botea, and A. F. Scott, “Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 8017–8020.
- [72] N. Barnes, P. Lieby, H. Dennet, J. Walker, C. McCarthy, N. Liu, and Y. Li, “Investigating the role of single-viewpoint depth data in visually-guided mobility,” *Journal of Vision*, vol. 11, no. 11, pp. 926–926, 2011.
- [73] C. McCarthy, N. Barnes, and P. Lieby, “Ground surface segmentation for navigation with a low resolution visual prosthesis,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4457–4460.
- [74] C. McCarthy and N. Barnes, “Time-to-contact maps for navigation with a low resolution visual prosthesis,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 2780–2783.
- [75] N. Parikh, L. Itti, M. Humayun, and J. Weiland, “Performance of visually guided tasks using simulated prosthetic vision and saliency-based cues,” *Journal of Neural Engineering*, vol. 10, no. 2, p. 026017, 2013.
- [76] M. Sanchez-Garcia, R. Martinez-Cantin, and J. J. Guerrero, “Indoor scenes understanding for visual prosthesis with fully convolutional networks.” in *VISIGRAPP (5: VISAPP)*, 2019, pp. 218–225.
- [77] M. Sanchez-Garca, R. Martinez-Cantin, and J. J. Guerrero, “Semantic and structural image segmentation for prosthetic vision,” *Plos one*, vol. 15, no. 1, p. e0227677, 2020.
- [78] M. Sanchez-Garcia, R. Martinez-Cantin, J. Bermudez-Cameo, and J. J. Guerrero, “Influence of field of view in visual prostheses design: Analysis with a vr system,” *Journal of Neural Engineering*, vol. 17, no. 5, p. 056002, 2020.
- [79] D. T. Hartong, E. L. Berson, and T. P. Dryja, “Retinitis pigmentosa,” *The Lancet*, vol. 368, no. 9549, pp. 1795–1809, 2006.
- [80] D.-Y. Yu and S. J. Cringle, “Retinal degeneration and local oxygen metabolism,” *Experimental eye research*, vol. 80, no. 6, pp. 745–751, 2005.

- [81] D. D. Zhou, J. D. Dorn, and R. 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.
- [82] D. L. Cheng, P. B. Greenberg, and D. A. Borton, "Advances in retinal prosthetic research: a systematic review of engineering and clinical characteristics of current prosthetic initiatives," *Current Eye Research*, vol. 42, no. 3, pp. 334–347, 2017.
- [83] N. H. Lovell, L. E. Hallum, S. Chen, S. Dokos, P. Byrnes-Preston, R. Green, L. Poole-Warren, T. Lehmann, and G. J. Suaning, "Advances in retinal neuroprosthetics," *Handbook of Neural Engineering*, pp. 337–356, 2007.
- [84] D. Nanduri, M. Humayun, R. Greenberg, M. McMahon, and J. Weiland, "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.
- [85] N. C. Sinclair, M. N. Shivdasani, T. Perera, L. N. Gillespie, H. J. McDermott, L. N. Ayton, and P. J. Blamey, "The appearance of phosphenes elicited using a suprachoroidal retinal prosthesis," *Investigative ophthalmology & visual science*, vol. 57, no. 11, pp. 4948–4961, 2016.
- [86] A. Barriga-Rivera, L. Bareket, J. Goding, U. A. Aregueta-Robles, and G. J. Suaning, "Visual prosthesis: interfacing stimulating electrodes with retinal neurons to restore vision," *Frontiers in neuroscience*, vol. 11, p. 620, 2017.
- [87] Y. Terasawa, A. Uehara, E. Yonezawa, T. Saitoh, K. Shodo, M. Ozawa, Y. Tano, and J. Ohta, "A visual prosthesis with 100 electrodes featuring wireless signals and wireless power transmission," *IEICE Electronics Express*, vol. 5, no. 15, pp. 574–580, 2008.
- [88] J. D. Weiland, W. Liu, and M. S. Humayun, "Retinal prosthesis," *Annu. Rev. Biomed. Eng.*, vol. 7, pp. 361–401, 2005.
- [89] T. T. Kien, T. Maul, and A. Bargiela, "A review of retinal prosthesis approaches," in *International Journal of Modern Physics: Conference Series*, vol. 9. World Scientific, 2012, pp. 209–231.
- [90] C. D. Eiber, N. H. Lovell, and G. J. Suaning, "Attaining higher resolution visual prosthetics: a review of the factors and limitations," *Journal of Neural Engineering*, vol. 10, no. 1, p. 011002, 2013.
- [91] S. Chen, L. Hallum, G. Suaning, and N. Lovell, "A quantitative analysis of head movement behaviour during visual acuity assessment under prosthetic vision simulation," *Journal of Neural Engineering*, vol. 4, no. 1, p. S108, 2007.
- [92] G. Dagnelie, "Visual prosthetics 2006: assessment and expectations," *Expert Review of Medical Devices*, vol. 3, no. 3, pp. 315–325, 2006.
- [93] M. S. Humayun, J. D. Weiland, G. Y. Fujii, R. Greenberg, R. Williamson, J. Little, B. Mech, V. Cimmarusti, G. Van Boemel, G. Dagnelie *et al.*, "Visual perception in a blind subject with a chronic microelectronic retinal prosthesis," *Vision research*, vol. 43, no. 24, pp. 2573–2581, 2003.

- [94] C. Choi, M. K. Choi, S. Liu, M. S. Kim, O. K. Park, C. Im, J. Kim, X. Qin, G. J. Lee, K. W. Cho *et al.*, “Human eye-inspired soft optoelectronic device using high-density mos 2-graphene curved image sensor array,” *Nature Communications*, vol. 8, no. 1, p. 1664, 2017.
- [95] J. S. Hayes, V. T. Yin, D. Piyathaisere, J. D. Weiland, M. S. Humayun, and G. Dagnelie, “Visually guided performance of simple tasks using simulated prosthetic vision,” *Artificial Organs*, vol. 27, no. 11, pp. 1016–1028, 2003.
- [96] C. Qiu, K. R. Lee, J.-H. Jung, R. Goldstein, and E. Peli, “Motion parallax improves object recognition in the presence of clutter in simulated prosthetic vision,” *Translational Vision Science & Technology*, vol. 7, no. 5, pp. 29–29, 2018.
- [97] F. Guo, Y. Yang, and Y. Gao, “Optimization of visual information presentation for visual prosthesis,” *International Journal of Biomedical Imaging*, vol. 2018, 2018.
- [98] B. Bourkiza, M. Vurro, A. Jeffries, and J. S. Pezaris, “Visual acuity of simulated thalamic visual prostheses in normally sighted humans,” *PloS one*, vol. 8, no. 9, p. e73592, 2013.
- [99] V. Vergnieux, M. J.-M. Macé, and C. Jouffrais, “Simplification of visual rendering in simulated prosthetic vision facilitates navigation,” *Artificial Organs*, vol. 41, no. 9, pp. 852–861, 2017.
- [100] —, “Wayfinding with simulated prosthetic vision: Performance comparison with regular and structure-enhanced renderings,” in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 2585–2588.
- [101] A. Perez-Yus, J. Bermudez-Cameo, G. Lopez-Nicolas, and J. J. Guerrero, “Depth and motion cues with phosphene patterns for prosthetic vision.” in *IEEE International Conference on Computer Vision.*, 2017, pp. 1516–1525.
- [102] H. Zhu, F. Meng, J. Cai, and S. Lu, “Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation,” *Journal of Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.
- [103] W. Wang, Y. Wang, Y. Wu, T. Lin, S. Li, and B. Chen, “Quantification of full left ventricular metrics via deep regression learning with contour-guidance,” *IEEE Access*, vol. 7, pp. 47 918–47 928, 2019.
- [104] Z. Zhang, C. Duan, T. Lin, S. Zhou, Y. Wang, and X. Gao, “Gvfom: a novel external force for active contour based image segmentation,” *Information Sciences*, vol. 506, pp. 1–18, 2020.
- [105] M. Wertheimer, “Laws of organization in perceptual forms.” *Psychologische Forschung*, pp. 301–350, 1938.
- [106] V. Bruce, P. R. Green, and M. A. Georgeson, *Visual perception: Physiology, psychology, & ecology*. Psychology Press, 2003.
- [107] D. D. Hoffman and M. Singh, “Saliency of visual parts,” *Cognition*, vol. 63, no. 1, pp. 29–78, 1997.

- [108] M. A. Minsky and S. Pappert, “Project mac progress report iv,” MIT Press, Cambridge, Mass., Tech. Rep., 1967.
- [109] C. R. Brice and C. Fennema, “Scene analysis using regions,” *Artificial Intelligence*, vol. 1, pp. 205–226, 01 1970.
- [110] R. M. Haralick and L. G. Shapiro, “Image segmentation techniques,” in *Applications of Artificial Intelligence II*, vol. 548. International Society for Optics and Photonics, 1985, pp. 2–10.
- [111] E. H. Adelson, “On seeing stuff: the perception of materials by humans and machines,” in *Human vision and electronic imaging VI*, vol. 4299. International Society for Optics and Photonics, 2001, pp. 1–12.
- [112] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” *arXiv preprint arXiv:1704.06857*, 2017.
- [113] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3150–3158.
- [114] M. Sanchez-Garcia, R. Martinez-Cantin, and J. J. Guerrero, “Indoor scenes understanding for visual prosthesis with fully convolutional networks,” in *14th International Conference on Computer Vision Theory and Applications, 2019*, pp 218-225, 2019.
- [115] M. Sun, B.-s. Kim, P. Kohli, and S. Savarese, “Relating things and stuff via object property interactions,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1370–1383, 2013.
- [116] J. Tighe, M. Niethammer, and S. Lazebnik, “Scene parsing with object instances and occlusion ordering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3748–3755.
- [117] J. Yao, S. Fidler, and R. Urtasun, “Describing the scene as a whole: joint object detection,” in *Proceedings of CVPR*. Citeseer, 2012.
- [118] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.
- [119] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, and J. J. Guerrero, “Panoroom: From the sphere to the 3d layout,” *arXiv preprint arXiv:1808.09879*, 2018.
- [120] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [121] A. Mallya and S. Lazebnik, “Learning informative edge maps for indoor scene layout prediction,” in *IEEE International Conference on Computer Vision*, 2015, pp. 936–944.

- [122] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [123] O. Ronneberger, P. Fischer, and T. 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.
- [124] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [125] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [126] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [127] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [128] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
- [129] J. M. Fácil, A. Concha, L. Montesano, and J. Civera, “Single-view and multi-view depth fusion,” *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 1994–2001, 2017.
- [130] N. Barnes, “The role of computer vision in prosthetic vision,” *Image and Vision Computing*, vol. 30, no. 8, pp. 478–479, 2012.
- [131] D. Feng and C. McCarthy, “Enhancing scene structure in prosthetic vision using iso-disparity contour perturbation maps,” in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 5283–5286.
- [132] M. Snaith, D. Lee, and P. Probert, “A low-cost system using sparse vision for navigation in the urban environment,” *Image and Vision Computing*, vol. 16, no. 4, pp. 225–233, 1998.
- [133] T. Sanocki, K. W. Bowyer, M. D. Heath, and S. Sarkar, “Are edges sufficient for object recognition?” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 1, p. 340, 1998.
- [134] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, “scikit-image: image processing in Python,” *PeerJ*, vol. 2, p. e453, 6 2014. [Online]. Available: <https://doi.org/10.7717/peerj.453>

- [135] J. L. Irons, T. Gradden, A. Zhang, X. He, N. Barnes, A. F. Scott, and E. McKone, "Face identity recognition in simulated prosthetic vision is poorer than previously reported and can be improved by caricaturing," *Vision research*, vol. 137, pp. 61–79, 2017.
- [136] Y. Lu, J. Wang, H. Wu, L. Li, X. Cao, and X. Chai, "Recognition of objects in simulated irregular phosphene maps for an epiretinal prosthesis," *Artificial organs*, vol. 38, no. 2, pp. E10–E20, 2014.
- [137] J. Hu, P. Xia, C. Gu, J. Qi, S. Li, and Y. Peng, "Recognition of similar objects using simulated prosthetic vision," *Artificial organs*, vol. 38, no. 2, pp. 159–167, 2014.
- [138] A. P. Fornos, J. Sommerhalder, B. Rappaz, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision, iii: Do the spatial or temporal characteristics of stimulus pixelization really matter?" *Investigative Ophthalmology & Visual Science*, vol. 46, no. 10, pp. 3906–3912, 2005.
- [139] H. Li, T. Han, J. Wang, Z. Lu, X. Cao, Y. Chen, L. Li, C. Zhou, and X. Chai, "A real-time image optimization strategy based on global saliency detection for artificial retinal prostheses," *Information Sciences*, vol. 415, pp. 1–18, 2017.
- [140] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.
- [141] M. Beyeler, A. Rokem, G. M. Boynton, and I. Fine, "Learning to see again: biological constraints on cortical plasticity and the implications for sight restoration technologies," *Journal of neural engineering*, vol. 14, no. 5, p. 051003, 2017.
- [142] H. Li, Y. Zeng, Z. Lu, X. Cao, X. Su, X. Sui, J. Wang, and X. Chai, "An optimized content-aware image retargeting method: toward expanding the perceived visual field of the high-density retinal prosthesis recipients," *Journal of neural engineering*, vol. 15, no. 2, p. 026025, 2018.
- [143] K. Cha, K. W. Horch, and R. A. Normann, "Mobility performance with a pixelized vision system," *Vision research*, vol. 32, no. 7, pp. 1367–1372, 1992.
- [144] G. S. Brindley and W. Lewin, "The sensations produced by electrical stimulation of the visual cortex," *The Journal of physiology*, vol. 196, no. 2, pp. 479–493, 1968.
- [145] X. Cao, H. Li, Z. Lu, X. Chai, and J. Wang, "Eye-hand coordination using two irregular phosphene maps in simulated prosthetic vision for retinal prostheses," in *IEEE 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2017, pp. 1–5.
- [146] C. McCarthy, D. Feng, and N. Barnes, "Augmenting intensity to enhance scene structure in prosthetic vision," in *IEEE International Conference on Multimedia and Expo workshops*, 2013, pp. 1–6.
- [147] W. Al-Atabany, M. Al Yaman, and P. Degenaar, "Extraspectral imaging for improving the perceived information presented in retinal prosthesis," *Journal of Healthcare Engineering*, 2018.

- [148] M. Boucart and T. H. C. Tran, "Object and scene recognition impairments in patients with macular degeneration," in *Object Recognition*. InTech, 2011.
- [149] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.
- [150] M. A. Peterson and G. Rhodes, *Perception of faces, objects, and scenes: Analytic and holistic processes*. Oxford University Press, 2003.
- [151] I. Biederman, "On the semantics of a glance at a scene," in *Perceptual Organization*. Routledge, 2017, pp. 213–253.
- [152] I. Biederman and G. Ju, "Surface versus edge-based determinants of visual recognition," *Cognitive Psychology*, vol. 20, no. 1, pp. 38–64, 1988.
- [153] J. J. Koenderink, "What does the occluding contour tell us about solid shape?" *Perception*, vol. 13, no. 3, pp. 321–330, 1984.
- [154] J. Wang, B. Cao, P. Yu, L. Sun, W. Bao, and X. Zhu, "Deep learning towards mobile applications," in *IEEE 38th International Conference on Distributed Computing Systems*, 2018, pp. 1385–1393.
- [155] M. P. H. Zapf, M.-Y. Boon, P. B. Matteucci, N. H. Lovell, and G. J. Suaning, "Towards an assistive peripheral visual prosthesis for long-term treatment of retinitis pigmentosa: evaluating mobility performance in immersive simulations," *Journal of neural engineering*, vol. 12, no. 3, p. 036001, 2015.
- [156] M. S. Humayun, J. D. Dorn, L. Da Cruz, G. Dagnelie, J.-A. Sahel, P. E. Stanga, A. V. Cideciyan, J. L. Duncan, D. Elliott, E. Filley *et al.*, "Interim results from the international trial of second sight's visual prosthesis," *Ophthalmology*, vol. 119, no. 4, pp. 779–788, 2012.
- [157] L. da Cruz, J. D. Dorn, M. S. Humayun, G. Dagnelie, J. Handa, P.-O. Barale, J.-A. Sahel, P. E. Stanga, F. Hafezi, A. B. Safran *et al.*, "Five-year safety and performance results from the Argus II retinal prosthesis system clinical trial," *Ophthalmology*, vol. 123, no. 10, pp. 2248–2254, 2016.
- [158] M. Humayun, "Preliminary results from argus ii feasibility study: a 60 electrode epiretinal prosthesis," *Investigative ophthalmology & visual science*, vol. 50, no. 13, pp. 4744–4744, 2009.
- [159] J. D. Dorn, A. K. Ahuja, A. Caspi, L. Da Cruz, G. Dagnelie, J.-A. Sahel, R. J. Greenberg, M. J. McMahon, Argus II Study Group *et al.*, "The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal prosthesis," *JAMA ophthalmology*, vol. 131, no. 2, pp. 183–189, 2013.
- [160] A. K. Ahuja, J. Dorn, A. Caspi, M. McMahon, G. Dagnelie, P. Stanga, M. Humayun, R. Greenberg, Argus II Study Group *et al.*, "Blind subjects implanted with the Argus II retinal prosthesis are able to improve performance in a spatial-motor task," *British Journal of Ophthalmology*, vol. 95, no. 4, pp. 539–543, 2011.

- [161] A. C. Ho, M. S. Humayun, J. D. Dorn, L. Da Cruz, G. Dagnelie, J. Handa, P.-O. Barale, J.-A. Sahel, P. E. Stanga, F. Hafezi *et al.*, “Long-term results from an epiretinal prosthesis to restore sight to the blind,” *Ophthalmology*, vol. 122, no. 8, pp. 1547–1554, 2015.
- [162] L. Da Cruz, B. F. Coley, J. Dorn, F. Merlini, E. Filley, P. Christopher, F. K. Chen, V. Wuyyuru, J. Sahel, P. Stanga *et al.*, “The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss,” *British Journal of Ophthalmology*, vol. 97, no. 5, pp. 632–636, 2013.
- [163] Y. H.-L. Luo, J. J. Zhong, and L. Da Cruz, “The use of Argus II retinal prosthesis by blind subjects to achieve localisation and prehension of objects in 3-dimensional space,” *Graefe’s Archive for Clinical and Experimental Ophthalmology*, vol. 253, no. 11, pp. 1907–1914, 2015.
- [164] G. Dagnelie, P. Christopher, A. Arditi, L. da Cruz, J. L. Duncan, A. C. Ho, L. C. Olmos de Koo, J.-A. Sahel, P. E. Stanga, G. Thumann *et al.*, “Performance of real-world functional vision tasks by blind subjects improves after implantation with the Argus II retinal prosthesis system,” *Clinical & experimental ophthalmology*, vol. 45, no. 2, pp. 152–159, 2017.
- [165] Y. H. Luo, J. J. Zhong, M. Clemo, and L. da Cruz, “Long-term repeatability and reproducibility of phosphene characteristics in chronically implanted Argus II retinal prosthesis subjects,” *American journal of ophthalmology*, vol. 170, pp. 100–109, 2016.
- [166] A. Kotecha, J. Zhong, D. Stewart, and L. da Cruz, “The Argus II prosthesis facilitates reaching and grasping tasks: a case series,” *BMC ophthalmology*, vol. 14, no. 1, p. 71, 2014.
- [167] M. S. Beauchamp, D. Oswalt, P. Sun, B. L. Foster, J. F. Magnotti, S. Niketeghad, N. Pouratian, W. H. Bosking, and D. Yoshor, “Dynamic stimulation of visual cortex produces form vision in sighted and blind humans,” *Cell*, vol. 181, no. 4, pp. 774–783, 2020.
- [168] L. Yue, J. D. Weiland, B. Roska, and M. S. Humayun, “Retinal stimulation strategies to restore vision: Fundamentals and systems,” *Progress in retinal and eye research*, vol. 53, pp. 21–47, 2016.
- [169] E. M. Barhorst-Cates, K. M. Rand, and S. H. Creem-Regehr, “The effects of restricted peripheral field-of-view on spatial learning while navigating,” *PloS one*, vol. 11, no. 10, p. e0163785, 2016.
- [170] S. Haymes, D. Guest, A. Heyes, and A. Johnston, “Mobility of people with retinitis pigmentosa as a function of vision and psychological variables.” *Optometry and vision science: official publication of the American Academy of Optometry*, vol. 73, no. 10, pp. 621–637, 1996.
- [171] T. Kuyk, J. L. Elliott, and P. S. Fuhr, “Visual correlates of obstacle avoidance in adults with low vision.” *Optometry and vision science: official publication of the American Academy of Optometry*, vol. 75, no. 3, pp. 174–182, 1998.

- [172] T. K. Lohmann, F. Haiss, K. Schaffrath, A.-C. Schnitzler, F. Waschkowski, C. Barz, A.-M. van der Meer, C. Werner, S. Johnen, T. Laube *et al.*, “The very large electrode array for retinal stimulation (vlars)—a concept study,” *Journal of neural engineering*, vol. 16, no. 6, p. 066031, 2019.
- [173] H. Ameri, T. Ratanapakorn, S. Ufer, H. Eckhardt, M. S. Humayun, and J. D. Weiland, “Toward a wide-field retinal prosthesis,” *Journal of neural engineering*, vol. 6, no. 3, p. 035002, 2009.
- [174] A. M. Alshaghthrah, “A study to develop a new clinical measure to assess visual awareness in tunnel vision,” Ph.D. dissertation, The University of Manchester (United Kingdom), 2014.
- [175] W. L. Kennedy, J. G. Rosten, L. M. Young, K. J. Ciuffreda, and M. I. Levin, “A field expander for patients with retinitis pigmentosa: a clinical study.” *American journal of optometry and physiological optics*, vol. 54, no. 11, pp. 744–755, 1977.
- [176] Y. He, N. T. Huang, A. Caspi, A. Roy, and S. R. Montezuma, “Trade-off between field-of-view and resolution in the thermal-integrated Argus II system,” *Translational vision science & technology*, vol. 8, no. 4, pp. 29–29, 2019.
- [177] A. P. Fornos, J. Sommerhalder, A. Pittard, A. B. Safran, and M. Pelizzone, “Simulation of artificial vision: Iv. visual information required to achieve simple pointing and manipulation tasks,” *Vision research*, vol. 48, no. 16, pp. 1705–1718, 2008.
- [178] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, “Recognizing scene viewpoint using panoramic place representation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2695–2702.
- [179] T. Endo, K. Hozumi, M. Hirota, H. Kanda, T. Morimoto, K. Nishida, and T. Fujikado, “The influence of visual field position induced by a retinal prosthesis simulator on mobility,” *Graefe’s Archive for Clinical and Experimental Ophthalmology*, vol. 257, no. 8, pp. 1765–1770, 2019.
- [180] S. A. Haymes, A. W. Johnston, and A. D. Heyes, “Relationship between vision impairment and ability to perform activities of daily living,” *Ophthalmic and Physiological Optics*, vol. 22, no. 2, pp. 79–91, 2002.
- [181] W. H. Dobelle, “Artificial vision for the blind by connecting a television camera to the visual cortex,” *ASAIO journal*, vol. 46, no. 1, pp. 3–9, 2000.
- [182] M. E. Brelén, F. Duret, B. Gérard, J. Delbeke, and C. Veraart, “Creating a meaningful visual perception in blind volunteers by optic nerve stimulation,” *Journal of neural engineering*, vol. 2, no. 1, p. S22, 2005.
- [183] C. Veraart, M.-C. Wanet-Defalque, B. Gérard, A. Vanlierde, and J. Delbeke, “Pattern recognition with the optic nerve visual prosthesis,” *Artificial organs*, vol. 27, no. 11, pp. 996–1004, 2003.
- [184] J. R. Ehrlich, L. V. Ojeda, D. Wicker, S. Day, A. Howson, V. Lakshminarayanan, and S. E. Moroi, “Head-mounted display technology for low-vision rehabilitation and vision enhancement,” *American journal of ophthalmology*, vol. 176, pp. 26–32, 2017.

- [185] K. Cha, K. Horch, and R. A. Normann, "Simulation of a phosphene-based visual field: visual acuity in a pixelized vision system," *Annals of biomedical engineering*, vol. 20, no. 4, pp. 439–449, 1992.
- [186] S. Cai, L. Fu, H. Zhang, G. Hu, and Z. Liang, "Prosthetic visual acuity in irregular phosphene arrays under two down-sampling schemes: a simulation study," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 5223–5226.
- [187] S. Chen, L. Hallum, G. Suaning, and N. Lovell, "Psychophysics of prosthetic vision: I. visual scanning and visual acuity," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 4400–4403.
- [188] H. C. Stronks and G. Dagnelie, "The functional performance of the argus ii retinal prosthesis," *Expert Review of Medical Devices*, vol. 11, no. 1, pp. 23–30, 2014.
- [189] J. Sommerhalder and A. P. Fornos, "Prospects and limitations of spatial resolution," in *Artificial Vision*. Springer, 2017, pp. 29–45.
- [190] K. Kihara and Y. Takeda, "Time course of the integration of spatial frequency-based information in natural scenes," *Vision research*, vol. 50, no. 21, pp. 2158–2162, 2010.
- [191] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.
- [192] A. Oliva, "Gist of the scene," in *Neurobiology of attention*. Elsevier, 2005, pp. 251–256.
- [193] A. Oliva and P. G. Schyns, "Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognitive psychology*, vol. 34, no. 1, pp. 72–107, 1997.
- [194] G. Denis, C. Jouffrais, C. Mailhes, and M. J. Macé, "Simulated prosthetic vision: improving text accessibility with retinal prostheses," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 1719–1722.
- [195] S. Chen, L. Hallum, N. Lovell, and G. J. Suaning, "Visual acuity measurement of prosthetic vision: a virtual-reality simulation study," *Journal of Neural Engineering*, vol. 2, no. 1, p. S135, 2005.
- [196] S. Nayak and R. K. Das, "Application of artificial intelligence (ai) in prosthetic and orthotic rehabilitation," in *Service Robotics*. IntechOpen, 2020.
- [197] C. Bühler, R. Hoelper, H. Hoyer, and W. Humann, "Autonomous robot technology for advanced wheelchair and robotic aids for people with disabilities," *Robotics and autonomous systems*, vol. 14, no. 2-3, pp. 213–222, 1995.
- [198] K.-H. Park, Z. Bien, J.-J. Lee, B. K. Kim, J.-T. Lim, J.-O. Kim, H. Lee, D. H. Stefanov, D.-J. Kim, J.-W. Jung *et al.*, "Robotic smart house to assist people with movement disabilities," *Autonomous Robots*, vol. 22, no. 2, pp. 183–198, 2007.

- [199] H. Uehara, H. Higa, and T. Soken, "A mobile robotic arm for people with severe disabilities," in *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*. IEEE, 2010, pp. 126–129.
- [200] C. Morrison, E. Cutrell, A. Dhareshwar, K. Doherty, A. Thieme, and A. Taylor, "Imagining artificial intelligence applications with people with visual disabilities using tactile ideation," in *Proceedings of the 19th international acm sigaccess conference on computers and accessibility*, 2017, pp. 81–90.
- [201] F. Dramas, S. J. Thorpe, and C. Jouffrais, "Artificial vision for the blind: a bio-inspired algorithm for objects and obstacles detection," *International Journal of Image and Graphics*, vol. 10, no. 04, pp. 531–544, 2010.
- [202] M. Macé, V. Guivarch, and C. Jouffrais, "Brain–computer interface for the blind: The benefits from a computer-based object recognition approach," *Perception ECVF abstract*, vol. 40, pp. 51–51, 2011.
- [203] C. A. Curcio, C. Owsley, and G. R. Jackson, "Spare the rods, save the cones in aging and age-related maculopathy," *Investigative ophthalmology & visual science*, vol. 41, no. 8, pp. 2015–2018, 2000.
- [204] V. Busskamp, J. Duebel, D. Balya, M. Fradot, T. J. Viney, S. Siegert, A. C. Groner, E. Cabuy, V. Forster, M. Seeliger *et al.*, "Genetic reactivation of cone photoreceptors restores visual responses in retinitis pigmentosa," *science*, vol. 329, no. 5990, pp. 413–417, 2010.
- [205] S. Richer, W. Stiles, L. Statkute, J. Pulido, J. Frankowski, D. Rudy, K. Pei, M. Tsipursky, and J. Nyland, "Double-masked, placebo-controlled, randomized trial of lutein and antioxidant supplementation in the intervention of atrophic age-related macular degeneration: the veterans last study (lutein antioxidant supplementation trial)," *Optometry-Journal of the American Optometric Association*, vol. 75, no. 4, pp. 216–229, 2004.
- [206] J. S. Heier, D. M. Brown, V. Chong, J.-F. Korobelnik, P. K. Kaiser, Q. D. Nguyen, B. Kirchhof, A. Ho, Y. Ogura, G. D. Yancopoulos *et al.*, "Intravitreal aflibercept (vegf trap-eye) in wet age-related macular degeneration," *Ophthalmology*, vol. 119, no. 12, pp. 2537–2548, 2012.
- [207] J. D. Weiland, A. K. Cho, and M. S. Humayun, "Retinal prostheses: current clinical results and future needs," *Ophthalmology*, vol. 118, no. 11, pp. 2227–2237, 2011.
- [208] Y. Li *et al.*, "Simulated artificial human vision: The effects of spatial resolution and frame rate on mobility," *Advances in Intelligent IT: Active Media Technology 2006*, vol. 138, p. 138, 2006.
- [209] A. Maeder *et al.*, "Mobility enhancement and assessment for a visual prosthesis," *SPIE Medical Imaging 2004: Physiology, Function, and Structure from Medical Images*, 2004.
- [210] V. Vergnienx, M. Macé, and C. Jouffrais, "Spatial navigation with a simulated prosthetic vision in a virtual environment," in *Conférence en Neurosciences Computationnelles (NeuroComp 2012)*, 2012.

- [211] C. McCarthy, J. G. Walker, P. Lieby, A. Scott, and N. Barnes, "Mobility and low contrast trip hazard avoidance using augmented depth," *Journal of neural engineering*, vol. 12, no. 1, p. 016003, 2014.
- [212] C. McCarthy and N. Barnes, "Surface extraction from iso-disparity contours," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 410–421.
- [213] D. Clark-Carter, A. Heyes, and C. Howarth, "The efficiency and walking speed of visually impaired people," *Ergonomics*, vol. 29, no. 6, pp. 779–789, 1986.
- [214] T. Meilinger, "The network of reference frames theory: A synthesis of graphs and cognitive maps," in *International conference on spatial cognition*. Springer, 2008, pp. 344–360.
- [215] J. A. Bagnell, D. Bradley, D. Silver, B. Sofman, and A. Stentz, "Learning for autonomous navigation," *IEEE Robotics & Automation Magazine*, vol. 17, no. 2, pp. 74–84, 2010.
- [216] R. C. Arkin and R. R. Murphy, "Autonomous navigation in a manufacturing environment." *IEEE Transactions on Robotics and Automation*, vol. 6, no. 4, pp. 445–454, 1990.
- [217] M. Piaggio, A. Sgorbissa, and R. Zaccaria, "Autonomous navigation and localization in service mobile robotics," in *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, vol. 4. IEEE, 2001, pp. 2024–2029.
- [218] D. Silver, J. A. Bagnell, and A. Stentz, "Active learning from demonstration for robust autonomous navigation," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 200–207.
- [219] J. Cheng, H. Cheng, M. Q.-H. Meng, and H. Zhang, "Autonomous navigation by mobile robots in human environments: A survey," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2018, pp. 1981–1986.
- [220] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.
- [221] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *The international journal of robotics research*, vol. 30, no. 7, pp. 846–894, 2011.
- [222] M. Hoy, A. S. Matveev, and A. V. Savkin, "Algorithms for collision-free navigation of mobile robots in complex cluttered environments: a survey," *Robotica*, vol. 33, no. 3, pp. 463–497, 2015.
- [223] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on intelligent vehicles*, vol. 1, no. 1, pp. 33–55, 2016.
- [224] M. Everett, Y. F. Chen, and J. P. How, "Motion planning among dynamic, decision-making agents with deep reinforcement learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3052–3059.

- [225] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online POMDP planning for autonomous driving in a crowd," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 454–460.
- [226] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [227] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.
- [228] T. Rabbani, F. Van Den Heuvel, and G. Vosselmann, "Segmentation of point clouds using smoothness constraint," *International archives of photogrammetry, remote sensing and spatial information sciences*, vol. 36, no. 5, pp. 248–253, 2006.
- [229] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3. IEEE, 2004, pp. 2149–2154.
- [230] R. F. Joseph and A. A. Godbole, "An intelligent traveling companion for visually impaired pedestrian," in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*. IEEE, 2014, pp. 283–288.
- [231] N. Kanwal, E. Bostanci, K. Currie, and A. F. Clark, "A navigation system for the visually impaired: a fusion of vision and depth sensor," *Applied bionics and biomechanics*, vol. 2015, 2015.
- [232] D. Nandini and K. Seeja, "A novel path planning algorithm for visually impaired people," *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 3, pp. 385–391, 2019.
- [233] S. Kammoun, F. Dramas, B. Oriolaand, and C. Jouffrais, "Route selection algorithm for blind pedestrian," in *ICCAS 2010*. IEEE, 2010, pp. 2223–2228.
- [234] M. Swobodzinski and M. Raubal, "An indoor routing algorithm for the blind: development and comparison to a routing algorithm for the sighted," *International Journal of Geographical Information Science*, vol. 23, no. 10, pp. 1315–1343, 2009.
- [235] S. M. LaValle and J. J. Kuffner Jr, "Randomized kinodynamic planning," *The international journal of robotics research*, vol. 20, no. 5, pp. 378–400, 2001.
- [236] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [237] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

- [238] J. Minguez and L. Montano, "Sensor-based robot motion generation in unknown, dynamic and troublesome scenarios," *Robotics and Autonomous Systems*, vol. 52, no. 4, pp. 290–311, 2005.
- [239] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Autonomous robot vehicles*. Springer, 1986, pp. 396–404.
- [240] J. Barraquand, L. Kavraki, J.-C. Latombe, R. Motwani, T.-Y. Li, and P. Raghavan, "A random sampling scheme for path planning," *The International Journal of Robotics Research*, vol. 16, no. 6, pp. 759–774, 1997.
- [241] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng *et al.*, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.
- [242] Y. Zhao, X. Geng, Q. Li, G. Jiang, Y. Gu, and X. Lv, "recognition of a virtual scene via simulated prosthetic vision," *Frontiers in Bioengineering and Biotechnology*, vol. 5, p. 58, 2017.
- [243] R. G. Golledge, "Path selection and route preference in human navigation: A progress report," in *International conference on spatial information theory*. Springer, 1995, pp. 207–222.
- [244] N. R. Council *et al.*, "Working group on mobility aids for the visually impaired and blind, electronic travel aids: New directions for research," 1986.
- [245] J. M. Benjamin, "The laser cane," *Bulletin of prosthetics research*, vol. 11, no. 2, pp. 443–450, 1974.
- [246] J. Armstrong, "Summary report of the research programme on electronic mobility aids," *Dep. of Psychology, Univ. of Nottingham, Nottingham, England*, 1973.
- [247] M. R. VanNewkirk, M. B. Nanjan, J. J. Wang, P. Mitchell, H. R. Taylor, and C. A. McCarty, "The prevalence of age-related maculopathy: the visual impairment project," *Ophthalmology*, vol. 107, no. 8, pp. 1593–1600, 2000.
- [248] J. R. Vingerling, I. Dielemans, A. Hofman, D. E. Grobbee, M. Hijmering, C. F. Kramer, and P. T. de Jong, "The prevalence of age-related maculopathy in the rotterdam study," *Ophthalmology*, vol. 102, no. 2, pp. 205–210, 1995.
- [249] J. D. Steinmetz, R. R. Bourne, P. S. Briant, S. R. Flaxman, H. R. Taylor, J. B. Jonas, A. A. Abdoli, W. A. Abrha, A. Abualhasan, E. G. Abu-Gharbieh *et al.*, "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study," *The Lancet Global Health*, vol. 9, no. 2, pp. e144–e160, 2021.
- [250] K. Stieger, C. Chauveau, and F. Rolling, "Preclinical studies on specific gene therapy for recessive retinal degenerative diseases," *Current gene therapy*, vol. 10, no. 5, pp. 389–403, 2010.

- [251] Y. Huang, V. Enzmann, and S. T. Ildstad, “Stem cell-based therapeutic applications in retinal degenerative diseases,” *Stem Cell Reviews and Reports*, vol. 7, no. 2, pp. 434–445, 2011.
- [252] I. Y.-H. Wong, M.-W. Poon, R. T.-W. Pang, Q. Lian, and D. Wong, “Promises of stem cell therapy for retinal degenerative diseases,” *Graefe’s Archive for Clinical and Experimental Ophthalmology*, vol. 249, no. 10, pp. 1439–1448, 2011.
- [253] W. A. Beltran, A. V. Cideciyan, A. S. Lewin, S. Iwabe, H. Khanna, A. Sumaroka, V. A. Chiodo, D. S. Fajardo, A. J. Román, W.-T. Deng *et al.*, “Gene therapy rescues photoreceptor blindness in dogs and paves the way for treating human x-linked retinitis pigmentosa,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 6, pp. 2132–2137, 2012.
- [254] V. K. Takahashi, J. T. Takiuti, R. Jauregui, and S. H. Tsang, “Gene therapy in inherited retinal degenerative diseases, a review,” *Ophthalmic genetics*, vol. 39, no. 5, pp. 560–568, 2018.
- [255] A. Farnum and G. Pelled, “New vision for visual prostheses,” *Frontiers in neuroscience*, vol. 14, p. 36, 2020.
- [256] H. Lorach, J. Wang, D. Y. Lee, R. Dalal, P. Huie, and D. Palanker, “Retinal safety of near infrared radiation in photovoltaic restoration of sight,” *Biomedical optics express*, vol. 7, no. 1, pp. 13–21, 2016.
- [257] F. Duret, M. E. Brelén, V. Lambert, B. Gérard, J. Delbeke, and C. Veraart, “Object localization, discrimination, and grasping with the optic nerve visual prosthesis,” *Restorative neurology and neuroscience*, vol. 24, no. 1, pp. 31–40, 2006.
- [258] Y. Lu, Y. Yan, X. Chai, Q. Ren, Y. Chen, and L. Li, “Electrical stimulation with a penetrating optic nerve electrode array elicits visuotopic cortical responses in cats,” *Journal of neural engineering*, vol. 10, no. 3, p. 036022, 2013.
- [259] V. Gaillet, A. Cutrone, F. Artoni, P. Vagni, A. M. Pratiwi, S. A. Romero, D. L. Di Paola, S. Micera, and D. Ghezzi, “Spatially selective activation of the visual cortex via intraneural stimulation of the optic nerve,” *Nature biomedical engineering*, vol. 4, no. 2, pp. 181–194, 2020.
- [260] N. J. Killian, M. Vurro, S. B. Keith, M. J. Kyada, and J. S. Pezaris, “Perceptual learning in a non-human primate model of artificial vision,” *Scientific reports*, vol. 6, no. 1, pp. 1–16, 2016.
- [261] E. Fernandez, F. Pelayo, S. Romero, M. Bongard, C. Marin, A. Alfaro, and L. Merabet, “Development of a cortical visual neuroprosthesis for the blind: the relevance of neuroplasticity,” *Journal of neural engineering*, vol. 2, no. 4, p. R1, 2005.
- [262] R. A. Normann, B. A. Greger, P. House, S. F. Romero, F. Pelayo, and E. Fernandez, “Toward the development of a cortically based visual neuroprosthesis,” *Journal of neural engineering*, vol. 6, no. 3, p. 035001, 2009.

- [263] S. R. Kane, S. F. Cogan, J. Ehrlich, T. D. Plante, D. B. McCreery, and P. R. Troyk, “Electrical performance of penetrating microelectrodes chronically implanted in cat cortex,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2153–2160, 2013.
- [264] R. A. Normann and E. Fernandez, “Clinical applications of penetrating neural interfaces and utah electrode array technologies,” *Journal of neural engineering*, vol. 13, no. 6, p. 061003, 2016.
- [265] S. Niketeghad, A. Muralidharan, U. Patel, J. D. Dorn, L. Bonelli, R. J. Greenberg, and N. Pouratian, “Phosphene perceptions and safety of chronic visual cortex stimulation in a blind subject,” *Journal of neurosurgery*, vol. 132, no. 6, pp. 2000–2007, 2019.
- [266] G. J. Chader, J. Weiland, and M. S. Humayun, “Artificial vision: needs, functioning, and testing of a retinal electronic prosthesis,” *Progress in Brain Research*, vol. 175, pp. 317–332, 2009.
- [267] J. Dowling, “Current and future prospects for optoelectronic retinal prostheses,” *Eye*, vol. 23, no. 10, pp. 1999–2005, 2009.
- [268] H. G. Sachs and V.-P. Gabel, “Retinal replacement—the development of microelectronic retinal prostheses—experience with subretinal implants and new aspects,” *Graefe’s archive for clinical and experimental ophthalmology*, vol. 242, no. 8, pp. 717–723, 2004.
- [269] S. Shim, K. Eom, J. Jeong, and S. J. Kim, “Retinal prosthetic approaches to enhance visual perception for blind patients,” *Micromachines*, vol. 11, no. 5, p. 535, 2020.
- [270] W. Tong, M. Stamp, N. V. Apollo, K. Ganesan, H. Meffin, S. Praver, D. J. Garrett, and M. R. Ibbotson, “Improved visual acuity using a retinal implant and an optimized stimulation strategy,” *Journal of neural engineering*, vol. 17, no. 1, p. 016018, 2019.
- [271] S. Rizzo, C. Belting, L. Cinelli, L. Allegrini, F. Genovesi-Ebert, F. Barca, and E. Di Bartolo, “The argus ii retinal prosthesis: 12-month outcomes from a single-study center,” *American journal of ophthalmology*, vol. 157, no. 6, pp. 1282–1290, 2014.
- [272] E. Zrenner, K. U. Bartz-Schmidt, H. Benav, D. Besch, A. Bruckmann, V.-P. Gabel, F. Gekeler, U. Greppmaier, A. Harscher, S. Kibbel *et al.*, “Subretinal electronic chips allow blind patients to read letters and combine them to words,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1711, pp. 1489–1497, 2011.
- [273] S. Chen, N. Lovell, and G. Suaning, “Effect on prosthetic vision visual acuity by filtering schemes, filter cut-off frequency and phosphene matrix: a virtual reality simulation,” in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2. IEEE, 2004, pp. 4201–4204.
- [274] J. T. Thorn, E. Migliorini, and D. Ghezzi, “Virtual reality simulation of epiretinal stimulation highlights the relevance of the visual angle in prosthetic vision,” *Journal of Neural Engineering*, vol. 17, no. 5, p. 056019, 2020.

- [275] M. Bach, M. Wilke, B. Wilhelm, E. Zrenner, and R. Wilke, "Basic quantitative assessment of visual performance in patients with very low vision," *Investigative ophthalmology & visual science*, vol. 51, no. 2, pp. 1255–1260, 2010.
- [276] M. Bach *et al.*, "The freiburg visual acuity test-automatic measurement of visual acuity," *Optometry and vision science*, vol. 73, no. 1, pp. 49–53, 1996.
- [277] M. S. Humayun, E. De Juan, G. Dagnelie, R. J. Greenberg, R. H. Propst, and D. H. Phillips, "Visual perception elicited by electrical stimulation of retina in blind humans," *Archives of ophthalmology*, vol. 114, no. 1, pp. 40–46, 1996.
- [278] P. Sinha, "Recognizing complex patterns," *nature neuroscience*, vol. 5, no. 11, pp. 1093–1097, 2002.
- [279] E. Fernández, A. Alfaro, and P. González-López, "Toward long-term communication with the brain in the blind by intracortical stimulation: Challenges and future prospects," *Frontiers in Neuroscience*, vol. 14, p. 681, 2020.
- [280] F. W. Campbell and J. G. Robson, "Application of fourier analysis to the visibility of gratings," *The Journal of physiology*, vol. 197, no. 3, p. 551, 1968.
- [281] C. Blakemore and F. W. Campbell, "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images," *The Journal of physiology*, vol. 203, no. 1, pp. 237–260, 1969.
- [282] A. P. Ginsburg and D. W. Evans, "Predicting visual illusions from filtered images based upon biological data (a)," *J. Opt. Soc. Am.*, vol. 69, page 1443, vol. 69, 1979.
- [283] L. Kauffmann, S. Ramanoël, and C. Peyrin, "The neural bases of spatial frequency processing during scene perception," *Frontiers in integrative neuroscience*, vol. 8, p. 37, 2014.
- [284] M. Baba, K. S. Sasaki, and I. Ohzawa, "Integration of multiple spatial frequency channels in disparity-sensitive neurons in the primary visual cortex," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 025–10 038, 2015.
- [285] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
- [286] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking deep belief network," *Frontiers in neuroscience*, vol. 7, p. 178, 2013.
- [287] S. Hussain, S.-C. Liu, and A. Basu, "Improved margin multi-class classification using dendritic neurons with morphological learning," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2014, pp. 2640–2643.
- [288] D. Neil and S.-C. Liu, "Minitaur, an event-driven fpga-based spiking network accelerator," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 12, pp. 2621–2628, 2014.

- [289] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations,” *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699–716, 2014.
- [290] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, “The spinnaker project,” *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.
- [291] T. Chauhan, T. Masquelier, A. Montlibert, and B. R. Cottureau, “Emergence of binocular disparity selectivity through hebbian learning,” *Journal of Neuroscience*, vol. 38, no. 44, pp. 9563–9578, 2018.
- [292] Q. Yu, H. Tang, K. C. Tan, and H. Li, “Rapid feedforward computation by temporal encoding and learning with spiking neurons,” *IEEE transactions on neural networks and learning systems*, vol. 24, no. 10, pp. 1539–1552, 2013.
- [293] D. Liu and S. Yue, “Visual pattern recognition using unsupervised spike timing dependent plasticity learning,” in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 285–292.
- [294] T. Masquelier and S. J. Thorpe, “Unsupervised learning of visual features through spike timing dependent plasticity,” *PLoS computational biology*, vol. 3, no. 2, p. e31, 2007.
- [295] A. Sboev, D. Vlasov, R. Rybka, and A. Serenko, “Solving a classification task by spiking neurons with stdp and temporal coding,” *Procedia computer science*, vol. 123, pp. 494–500, 2018.
- [296] P. U. Diehl and M. Cook, “Unsupervised learning of digit recognition using spike-timing-dependent plasticity,” *Frontiers in computational neuroscience*, vol. 9, p. 99, 2015.
- [297] E. M. Izhikevich, “Simple model of spiking neurons,” *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1569–1572, 2003.
- [298] J. Guckenheimer and J. Labouriau, “Bifurcation of the hodgkin and huxley equations: a new twist,” *Bulletin of Mathematical Biology*, vol. 55, no. 5, p. 937, 1993.
- [299] Ş. Mihalas and E. Niebur, “A generalized linear integrate-and-fire neural model produces diverse spiking behaviors,” *Neural computation*, vol. 21, no. 3, pp. 704–718, 2009.
- [300] W. Gerstner and R. Naud, “How good are neuron models?” *Science*, vol. 326, no. 5951, pp. 379–380, 2009.
- [301] Y.-H. Liu and X.-J. Wang, “Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron,” *Journal of computational neuroscience*, vol. 10, no. 1, pp. 25–45, 2001.
- [302] H. Markram, W. Gerstner, and P. J. Sjöström, “A history of spike-timing-dependent plasticity,” *Frontiers in synaptic neuroscience*, vol. 3, p. 4, 2011.

- [303] Y. LeCun, “The mnist database of handwritten digits,” <http://yann.lecun.com/exdb/mnist/>, 1998.
- [304] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [305] A. Delorme, L. Perrinet, and S. J. Thorpe, “Networks of integrate-and-fire neurons using rank order coding b: Spike timing dependent plasticity and emergence of orientation selectivity,” *Neurocomputing*, vol. 38, pp. 539–545, 2001.
- [306] S. Celebrini, S. Thorpe, Y. Trotter, and M. Imbert, “Dynamics of orientation coding in area v1 of the awake primate,” *Visual neuroscience*, vol. 10, no. 5, pp. 811–825, 1993.
- [307] T. J. Gawne, T. W. Kjaer, and B. J. Richmond, “Latency: another potential code for feature binding in striate cortex,” *Journal of neurophysiology*, vol. 76, no. 2, pp. 1356–1360, 1996.
- [308] W. Maass, “Neural computation with winner-take-all as the only nonlinear operation,” in *Advances in neural information processing systems*, 2000, pp. 293–299.
- [309] D. Ferster and S. Lindström, “An intracellular analysis of geniculo-cortical connectivity in area 17 of the cat.” *The Journal of physiology*, vol. 342, no. 1, pp. 181–215, 1983.
- [310] K. Tanaka, “Organization of geniculate inputs to visual cortical cells in the cat,” *Vision research*, vol. 25, no. 3, pp. 357–364, 1985.
- [311] R. Gütig, R. Aharonov, S. Rotter, and H. Sompolinsky, “Learning input correlations through nonlinear temporally asymmetric hebbian plasticity,” *Journal of Neuroscience*, vol. 23, no. 9, pp. 3697–3714, 2003.
- [312] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual neuroscience*, vol. 9, no. 2, pp. 181–197, 1992.
- [313] M. Carandini and D. J. Heeger, “Normalization as a canonical neural computation,” *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.
- [314] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [315] J. Chu, A. GuoLu, L. Wang, C. Pan, and S. Xiang, “Indoor frame recovering via line segments refinement and voting,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 1996–2000.
- [316] H. Guo, Y. Wang, Y. Yang, S. Tong, Y. Zhu, and Y. Qiu, “Object recognition under distorted prosthetic vision,” *Artificial Organs*, vol. 34, no. 10, pp. 846–856, 2010.

Appendix

Appendix for Schematic representation of indoor scenes

Phosphene generation

In order to simulate the visual perception in prosthetic vision, our phosphenes generated are idealized representations of the percepts feasible in the current implants. Concretely, our phosphene map configuration is similar to the framework described in [48]. First, once the image has been processed according to the chosen processing method, the image is sampled at each location in a given phosphene layout. The most commonly phosphene patterns adopted by SPV in the literature [45] are square or hexagonal array. Due to the prevalence of vertical lines in indoor scenarios [315] we opted to use a square pattern (Fig A0.1 (middle)). The simulation of indoor images with square pattern keeps the structure of each line as it really is, while the hexagonal pattern tends to distort it due to the deviation of the phosphenes when forming hexagons (Fig A0.1 (right)).



Figure A0.1: Vertical lines in indoor environments. (a) Example of indoor scene, (b) prevalence of vertical lines, (c) square pattern and (d) hexagonal pattern. As a result of the abundance of vertical lines in indoor environments, we opted to use the square grid.

Similarly to many SPV studies [45], phosphenes are approximated as grayscale circular dots with a Gaussian luminance profile. The luminance profile of each phosphene has maximum intensity at the center and gradually decays to the periphery, following an

unnormalized Gaussian function $G(x, y)$ defined in Eq (2.1).

$$G(x, y) \propto \exp \left\{ \frac{(x - \mu_x)^2 + (y - \mu_y)^2}{2\sigma^2 lum} \right\} \quad (2.1)$$

The intensity of a phosphene is a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ of the intensity of the pixels covering the same location in the processed image. Usually, f is the mean function, but we found that the median function preserved the line structures of SIE-OMS and Edges. In addition, each sampled pixel intensity (i) is quantified to each individual phosphene's dynamic range as:

$$\mathcal{L}(x, y) = \frac{\mathcal{I}(x, y)}{\max(l)} \quad (2.2)$$

In Eq (2.2), l is the number of gray levels intensity of the phosphene. The size and brightness are directly proportional to the quantified sampled pixel intensities. Then, the phosphene profile $P(x, y)$ is applied to every phosphene resulting in the final image:

$$P(x, y) = lum \cdot G(x, y) \quad (2.3)$$

where $P(x, y)$ represents the pixel value at the coordinates (x, y) of the stimulus image (Eq (2.3)).

Furthermore, several important works have demonstrated that exists a degree of incomplete phosphene visual field maps (dropout) in implanted patients. This is because of the high threshold required values to elicit phosphenes that placed in areas with a high proportion of dead nerve cells [93]. This results in a lower resolution than the number of electrode elements. We included a 10% dropout of phosphenes. Those phosphenes were initially selected randomly but kept the same for all the images. Note that our prosthetic vision simulation does not integrate all aspects of visual appearance reported by implanted subjects (e.g. distortion effect [145, 316]).

Appendix for Influence of field of view

Spherical image projection in the VR system

Since the participants only rotate the head and body during the experiment, in our SPV system, we only consider rotation movements. In the absence of translation, all the visual information of the environment can be encoded in a spherical image, i.e. an equirectangular panorama. Thus, if the rotation of the head is known, the image received by the subject through the screen can be virtually simulated from a panoramic image.

To achieve the effect of the stimulus image on the VR system we combine two different projections models: the spherical projection that models the projection on the panoramic image and the perspective projection that models the projection on the viewfinder screen.

In the Oculus VR system, the inertial measurement unit (IMU) collect the information of the head movement made by the user in form of quaternions like $q = a + bi + cj + dk$. This information is used to generate the rotation matrix which describes the basis of the absolute reference system used by the IMU with respect to the screen reference system as:

$$R_{imu} = \begin{pmatrix} 1 - b^2 - c^2 & 2ab - 2cd & 2ac + 2bd \\ 2ab + 2cd & 1 - 2a^2 - 2c^2 & 2cb + 2ad \\ 2ac - 2bd & 2bc + 2ad & 1 - 2a^2 - 2b^2 \end{pmatrix} \quad (3.4)$$

This rotation matrix defines the head orientation and, therefore, it can be used to find the area of the panoramic scene that is currently being observed. Then, the selected area is projected on the screen.

For the generation of the screen image, we calculate the ray corresponding to each phosphene on the screen as

$$\vec{v} = R_{pan}^T R_{imu}^T K^{-1} u \quad (3.5)$$

where $\vec{v} = (v_x, v_y, v_z)^T$ is the direction vector of the ray in the panorama reference system, R_{imu} is the rotation matrix from the IMU data, R_{pan} is the the basis of the panorama reference system with respect to the absolute reference system of the IMU and $u = (j, i, 1)^T$ are the coordinates of each phosphene on the screen image. K is the calibration matrix used in the perspective projection that models the projection in the screen.

The spherical coordinates of the ray \vec{v} , known as azimuth ϕ and elevation θ angles, and computed as:

$$\phi = \text{atan2}(v_y, v_x) \quad (3.6)$$

$$\theta = \sin^{-1} \frac{v_z}{\sqrt{v_x^2 + v_y^2 + v_z^2}} \quad (3.7)$$

and are related with the coordinates of the pixels in the panorama though the linear mapping,

$$x_{pan} = x_0 + \frac{\phi}{2\pi}W \quad (3.8)$$

$$y_{pan} = y_0 + \frac{\theta}{\pi}H \quad (3.9)$$

where (x_0, y_0) is the center of the panoramic image, and (W, H) are, respectively, the width and height of the panoramic image in pixels.