Media Engineering and Technology Faculty
German University in Cairo

Eye Detection and Face Recognition for Visual Prostheses

Bachelor Thesis

Author:	Hesham Mohamed Moneer
Supervisors:	Dr. Seif Eldawlatly
	Msc. Eng. Reham Elnabawy
Submission Date:	12 June, 2022

Media Engineering and Technology Faculty
German University in Cairo

Eye Detection and Face Recognition for Visual Prostheses

Bachelor Thesis

Author:	Hesham Mohamed Moneer
Supervisors:	Dr. Seif Eldawlatly
	Msc. Eng. Reham Elnabawy
Submission Date:	12 June, 2022

This is to certify that:

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

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

---

Hesham Mohamed Moneer
12 June, 2022

Acknowledgments

I would like to thank everyone who contributed to the work described in this thesis. First, I would love to thank my supervisors Dr. Seif Eldawlatly and Eng. Reham Elnabawy for their continual support and instructions. Second, I want to thank all my friends and family members who volunteered to undergo the experiments of this thesis.

Abstract

Blindness is one of the most common disabilities worldwide. It imposes large personal and societal costs. Although advancements in surgery, medication, and gene therapies offer great treatment options, blindness caused by severe damage in the retina, the optic nerve, or the brain might not be effectively treated with these offered options. A possible last resort to such cases is **Visual Prostheses**. **Visual Prostheses** are implantable medical devices that can provide a very limited vision with relatively low **Spatial Resolution** to these patients. However, due to the low resolution offered, the ability of the implantees to do everyday activities, like reading, crossing the road, or recognizing the identity of a seen face is still very limited. The aim of this thesis is to help these implantees in recognizing faces and facial details. This is done through processing the captured real-time scene before being transmitted with lower resolution to the brain to make the faces more recognizable. The processing is to be done using machine learning and computer vision techniques, and it is to be applied and tested in the form of virtual–reality visual models (simulations) of prosthetic vision administered to normally sighted subjects. The enhancement methods explored were applying Histogram equalization (i.e. enhancing contrast) to the whole visual field or applying face-specific Histogram equalization, magnification of the face to fill the visual field with it, caricaturing the face (i.e. exaggerating facial features, e.g. making a small nose smaller, thick lips thicker, etc.), highlighting the face emotion, or focusing on the face of the talking person. Two experiments were conducted: one on a computer screen and the other using a head-mounted display (HMD). The conducted experiments show that filling the visual field with a detected face has significantly increased the subjects ability in recognizing faces. Therefore, it was concluded that a combination of visual field histogram equalization, magnification of the face of interest to fill the visual field, and applying caricaturing to that face all together make the most promising enhancement approach.

Contents

Acknowledgments **V**

Abstract **VI**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Thesis Aim . . . 1
1.3 Thesis Outcomes . . . 2
1.4 Thesis Organization . . . 2

2 Background **3**
2.1 Visual Prostheses . . . 3
2.2 Simulations of Prosthetic Vision (SPVs) . . . 7
2.3 Facial detection . . . 11
2.4 Face Landmark Detection . . . 14
2.5 Face Caricaturing . . . 14
2.6 Related Work . . . 16
2.6.1 Histogram Equalization . . . 16
2.6.2 Region of interest magnification . . . 17
2.6.3 Caricaturing . . . 18

3 Methodologies **20**
3.1 Simulation Configuration . . . 21
3.2 Video Simulation . . . 21
3.2.1 Update Bounding Boxes . . . 22
3.2.2 Apply Bounding Boxes . . . 22
3.3 Singleton Image Simulation . . . 23
3.3.1 Image Preprocessing . . . 24
3.3.2 Drawing Phosphenes . . . 25
3.4 Enhancement Modules . . . 27
3.4.1 Face-specific histogram equalization (FSHE) . . . 28
3.4.2 Emotion Recognition (ER) . . . 29
3.4.3 Talking Detection (TD) . . . 30
3.4.4 Caricaturing . . . 31

3.5 Supported Modes . . . 34
3.5.1 Simulation Mode (*simode*) . . . 34
3.5.2 Enhancement Mode (*facesMode*) . . . 35
3.6 Computer Screen Experiment . . . 37
3.6.1 Experiment Groups . . . 37
3.6.2 Experiment Tests . . . 39
3.7 Virtual Reality Experiment . . . 41
3.7.1 Experiment Groups . . . 42
3.7.2 Experiment Tests . . . 44

4 Results **46**
4.1 Simulation output images . . . 46
4.2 Computer Screen Experiment . . . 47
4.3 Virtual Reality Experiment . . . 52
4.4 Applicability in real-time . . . 56

5 Conclusion and Future Work **57**
5.1 Conclusion . . . 57
5.2 Limitations and Future Work . . . 58

Appendix **60**

A Lists **61**
List of Abbreviations . . . 61
List of Figures . . . 64

References **66**

Chapter 1

Introduction

1.1 Motivation

Visual prostheses can help restore vision for patients with acquired blindness (as opposed to congenital blindness). It works by electrically stimulating functional visual structures using an electrode array. Engineering challenges, however, limit the maximum number of implantable electrodes and the size of the implant. Therefore, vision restored by visual prostheses suffer from low spatial resolution.

In a group conversation setting, it is very valuable for a group member to do the following efficiently:

1. recognize the other group members,
2. distinguish alike faces,
3. identify the talking person(s)
4. see the facial expressions of the talking person because facial expressions convey emotions that words cannot portray.

The efficiency of doing these tasks in this setting for a visual prostheses implantee is very low when compared to a normally sighted person due to that prostheses offered low resolution. Facial recognition, especially, is an important task in everyday activities, and a limited recognition ability will affect all these activities negatively. Therefore, it is very crucial for the implantees to do these mentioned tasks more efficiently.

1.2 Thesis Aim

The aim of this thesis is to improve visual prostheses implantees efficiency in the tasks mentioned in section 1.1 using Machine Learning (ML) models and Computer Vision (CV) techniques. This will in turn facilitate many everyday activities for these patients.

1.3 Thesis Outcomes

In this thesis, many approaches were explored to improve the implantees efficiency in doing the tasks mentioned in section 1.1. In all approaches, the image contrast was increased, to overcome any unclarities caused by poor lighting conditions, and it was proceeded with further processing on the face region. The explored approaches were the following:

1. Viola Jones Face Magnification (VJm): a face of interest was detected and magnified to fill the visual field.
2. Viola Jones Face Caricaturing (VJc): a face of interest was magnified to fill the visual field with caricaturing applied.
3. Viola Jones Talking Detection (VJtd): the talking face was detected and was magnified to fill the visual field.
4. Viola Jones Emotion Recognition (VJer): a face of interest was magnified to fill the visual field, and the emotion expressed by that face was highlighted.
5. Viola Jones Face-specific Histogram Equalization (VJhe): a face of interest was magnified to fill the visual field, and the contrast of that face was increased.

The approaches were applied on normally sighted subjects in the form of simulations. The efficiency of the subjects to do the tasks mentioned in section 1.1 was assessed. Efficiency was quantified in the form of accuracy (e.g. the percentage of the faces correctly recognized by the subject), confidence level (i.e. how confident a subject was of their recognition on a scale of 1 to 5), and response time (i.e. how long did it take a subject to recognize a face).

The conducted experiments show that VJc is the best approach taking the following into consideration: subjects' accuracy in recognizing faces, subjects' response time, subjects' confidence level, subjects' accuracy in distinguishing similar or different faces, subjects' accuracy in detecting the talking person, subjects' ability to capture facial expressions, and algorithm's complexity.

1.4 Thesis Organization

The thesis is divided into four main chapters. First, the Background (chapter 2) that provides a walk-through over the main concepts of the thesis and summarizes what was found in the literature of related work to visual prostheses, Simulations of Prosthetic Vision (SPVs), facial detection in CV, face landmark detection, and face caricaturing. Second, the thesis discusses Methodologies (chapter 3) that gives an overview of the implemented simulation and the conducted experiments. Third, the thesis outlines the Results (chapter 4) of both the implemented simulation and the conducted experiments. At last, the thesis discusses (chapter 5) the Conclusion and recommendations for Future Work.

Chapter 2

Background

2.1 Visual Prostheses

Millions of patients suffer from vision loss, or are gradually losing vision, due to retinal degenerative diseases such as retinitis pigmentosa (RP) or age-related macular degeneration (AMD) or because of accidents or injuries [1]. Vision restoration has a long history in biomedical engineering. An alternative approach to restore vision, which is discussed in this thesis, would be using implantable devices called Visual Prostheses.

Past studies have shown that there is a relationship between electrical stimulation of parts of the Visual Pathway and visual sensation, and so an artificial vision can be electrically stimulated. This is the core of Visual Prostheses: for blindness caused by interruption of the normal flow of signals along the Visual Pathway, exciting the neurons beyond the damaged site can induce perception of Phosphenes (see Figure 2.1) [2].

Figure 2.1: Phosphenes are visual sensations caused by means other than light stimulation. Stimulation could be mechanical, magnetic, electrical, ionizing radiation, etc. A phosphene in this thesis will refer to an electrically stimulated spot of light in the visual field [3]. The figure shows the possible perception of phosphenes generated by stimulating 4 electrodes simultaneously [2].

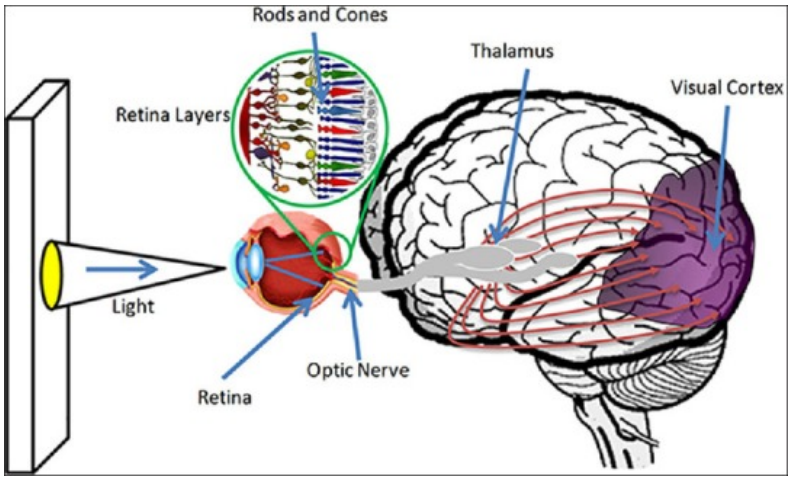


Figure 2.2: The Visual Pathway is the anatomical structures responsible for the conversion of light into electrical action potentials that can be interpreted by the brain. The most important structures of them for this thesis are the following: *1. Retina, 2. Optic Nerve, 3. Lateral Geniculate Nucleus (LGN) of the Thalamus, 4. Visual Cortex* because Visual Prostheses implants can be in direct contact with one of these structures [1].

Stimulating the Visual Pathway (see Figure 2.2) to produce visual sensation was first described by Franklin Cavallo and LeRoy. In 1755, LeRoy produced visual sensation by electrically stimulating the eye of a blind man. Later in 1929, a German neurosurgeon observed that a patient saw a spot of light when an electrical current was used to simulate his visual cortex [1].

There are several approaches to stimulate prosthetic vision. The stimulated visual structure could be the retina, the optic nerve, the lateral geniculate nucleus (LGN), or the visual cortex, as shown in Figure 2.3. The microelectronic device should directly contact functional neural elements. After implantation, implantees restore some degree of vision. The implantees description of the Phosphenes is depicted on Figure 2.1. Learning and rehabilitation strategies are necessary to make use of the Neuroplasticity of the visual system. This will lead to an improving correlation between the physical world and evoked phosphenes as shown in Figure 2.4.

In order to develop fully functional visual prosthesis, a multidisciplinary team needs to be formed. The team shall comprise engineers (to design circuits, systems, microstructures, and signal processing techniques) and specialists in the clinical and surgical fields (to surgically implant the device, monitor subjects health pre- and post-implantation, and to design clinical tests to assess the efficiency of the implant) closely working together (see Figure 2.5).

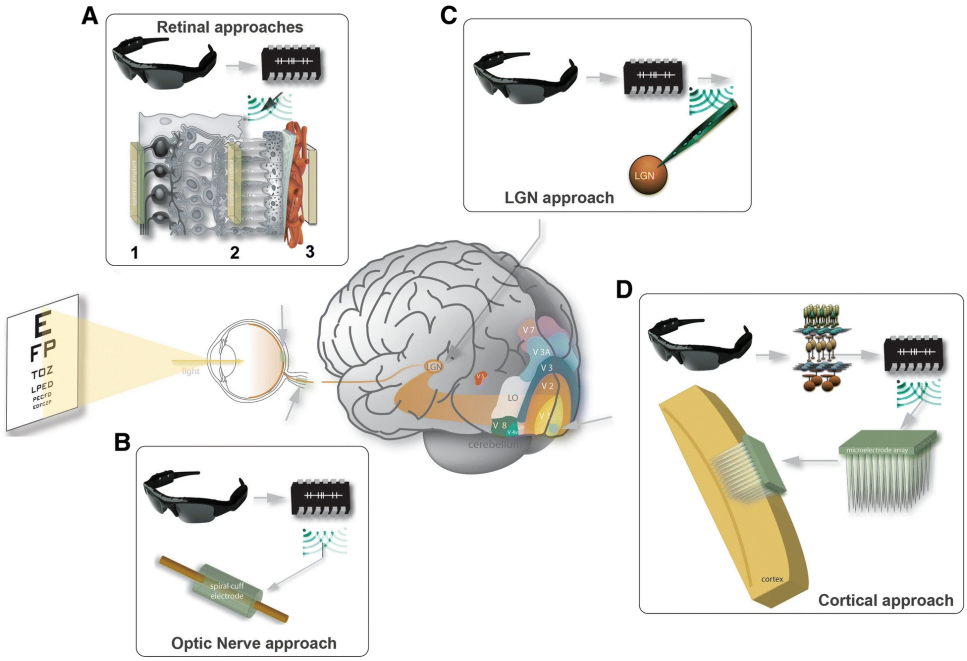


Figure 2.3: Main approaches of visual prosthesis.
A. Cross-section of the retina. Retinal implants approaches are:
1. *Epiretinal*, 2. *Subretinal* and 3. *Suprachordial*.
B. Optic nerve approach.
C. LGN approach.
D. Cortical approach. [2]

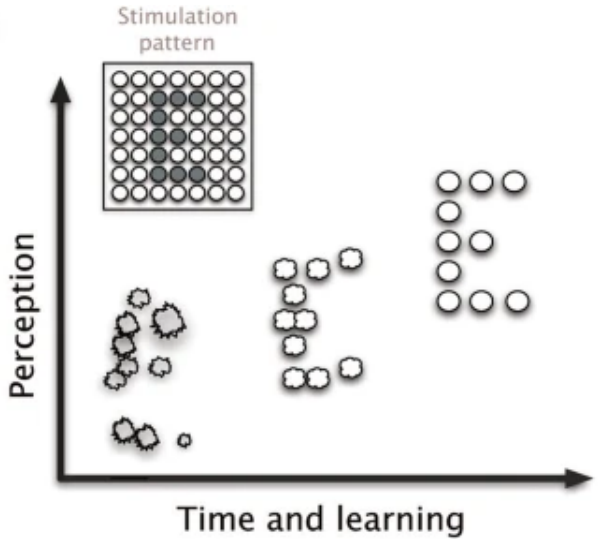


Figure 2.4: The neuroplasticity is the ability of the brain to rewire itself in order to develop or recover from injury. The figure shows neuroplasticity improving, with suitable learning and rehabilitation strategies, the poor perception of the letter **E** over time [2].

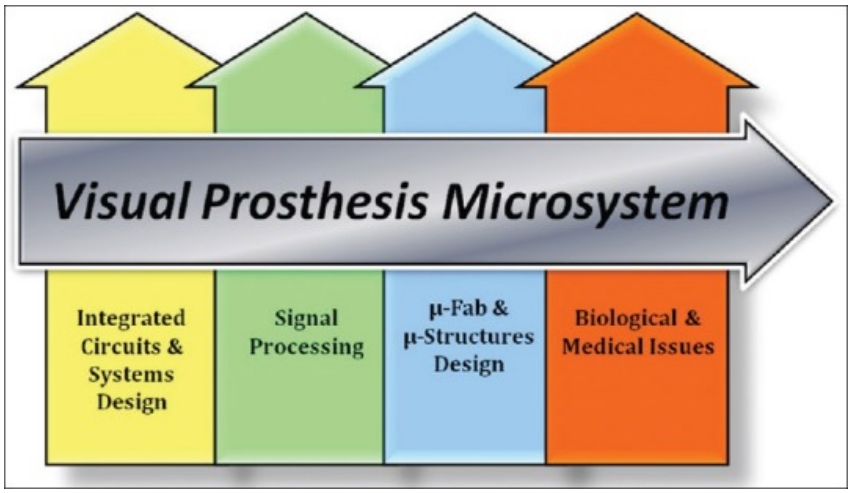


Figure 2.5: Visual Prosthesis Design Thrusts [1].

In general, all Visual Prostheses approaches share a common set of components:

1. Outside the body, and it contains
   (a) a camera mounted on standard glasses,
   (b) a micro-processing unit that processes and transforms the visual scene into electrical pulses,
   (c) a transmitter that transmits the pulses as radio-frequency waves to the implanted device,
   (d) and a power supply.
2. Implanted device: an electrode array implanted in the visual pathways near the target neurons

A visual acuity of 20/40 means that the person can see the same amount of detail from 20 feet away as the average person would see from 40 feet away [4]. The Visual Acuity offered by visual neuroprostheses is very low: best records were scored by ARGUS II and Alpha IMS to be 20/1260 and 20/546 respectively. This happens due to the low number of electrodes compared to the number of receptor cells: increasing the number of the electrodes causes crosstalk and interference problems. The current approach, therefore, is to equip the micro-processing unit with image processing techniques to facilitate facial recognition, reading, and other daily tasks [2].

2.2 Simulations of Prosthetic Vision (SPVs)

Most of the past studies, and this thesis as well, did not have access to actual visual prostheses implantees. Therefore, an alternative to having actual implantees was to proceed in the studies with simulations of visual prostheses applied to normally sighted subjects. Moreover, SPV based psychological studies are very crucial when it comes to assessing multiple things like: what is the number of phosphenes, angular resolution, processing techniques, and rehabilitation methods needed to allow implantees to do everyday activities efficiently? Therefore, the SPV must be an immersive experience to improve realism and effectiveness of the simulation. The simulation is done through head-mounted displays (HMDs) that act like tiny computer screens displaying the simulated image. Also, a head-mounted video camera is required to capture the real-time scene around the subject compatible with their head motion. In addition, the compatibility of the phosphene field with the eye motion can be achieved through an eye tracker.

A single phosphene in the visual field could be thought of as a single pixel in a raster image. Taking this idea of phosphenes as the building blocks of the visual field, investigators managed to simulate prosthetic vision based on the descriptions of singular phosphenes from the literature. In the future, prosthetic vision is expected to be provided in this form where a single phosphene constitutes a building block of the visual field.

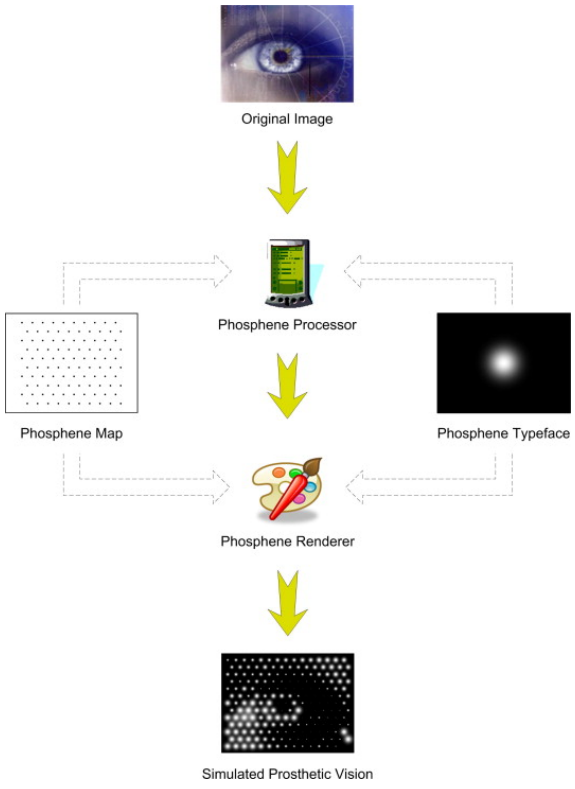


Figure 2.6: The process of simulating prosthetic vision [5].

As shown in Figure 2.6, an SPV can be divided into the following modules:

1. **Phosphene typeface**: phosphene style having varying size, luminance, color, and other attributes. These attribute indices are referred to as **Phosphene Modulation Indices** or PMIs.
2. **Phosphene map**: phosphene location in the visual field.
3. **Phosphene processor**: it takes a visual scene, and for each phosphene in the phosphene map, calculates the appropriate PMI. It also might take the phosphene typeface into consideration.
4. **Phosphene renderer**: it displays the phosphene field constructed by the phosphene processor.

The simulation process, as shown in Figure 2.6, starts with choosing the **phosphene map** and the **phosphene typeface** for each phosphene. A single typeface is typically used for all phosphenes in the same study. The processor then uses the typeface and the map to translate images into PMIs and sends them to the renderer.

The SPVs in the literature [5] adopted different PMIs and phosphene maps that are summarized below:

- Phosphene shape: [Circle, Square, Gaussian]
  The Gaussian Distribution (shown in Figure 2.7) in general is not bounded; it continues to infinity. However, Gaussian profiles in past studies are intentionally truncated to lower the memory footprint and to speedup processing time.

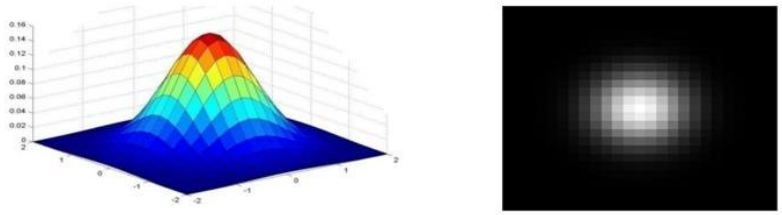


Figure 2.7: Gaussian Distribution shape [6].

- Phosphene color: gray
  Colored phosphenes have been observed by patients (blue, orange, yellow, or multicolored). However, colorless (white) phosphenes are the most dominant in the literature. In addition, no systematic correlation between electrical stimulus parameters and the phosphene color was found. Also, the same subject would often report the same color. Therefore, similar-colored gray phosphenes are adopted by every SPV in the literature.

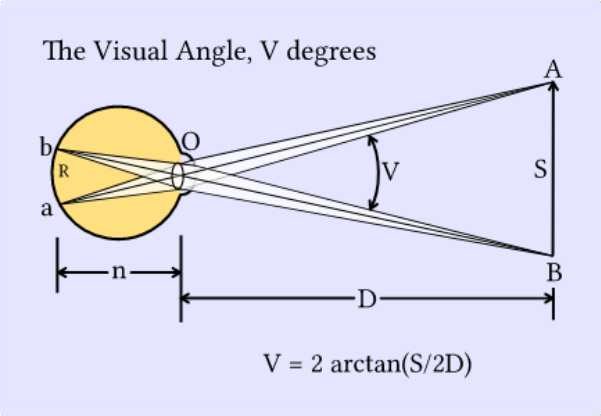


Figure 2.8: A depiction of the visual angle [7].

- Phosphene size: $[0.4° - 2.0°]$ or $[16px - 80px]$
  Patients reported phosphene sizes ranging from $5mm$ to $20mm$ at arm's length ($in\ this\ context,\ arm's\ length \approx 57.3cm \therefore [5mm{-}20mm]\ at\ arm's\ length \equiv [0.5°{-}2.0°]$)
  In the literature, the size is reported in Visual Angle that is the angle a viewed object subtends at the eye. (see Figure 2.8). Calculating what that range is equivalent to on a $96 \times 96\,dpi$ laptop screen:

$$
\begin{aligned}
\because 1in &= 25.4mm \\
\therefore 1mm &\equiv \frac{96dpi}{25.4mm} \cong 4px \\
\because 1mm\ at\ arm's\ length &= 0.1°\ visual\ angle \\
\therefore 1° &\equiv 40px\ at\ arm's\ length\ (\approx 57.3cm) \\
\therefore [0.4°{-}2.0°] &\equiv [16px{-}80px](at\ arm's\ length)
\end{aligned}
\tag{2.1}
$$

- Visual field lattice: [Square, bitmap, irregular, hexagonal, stochastic, log-polar]
  Most reported phosphene maps by implantees are of irregular shape. Most of the adopted lattices in SPVs, however, are either square or hexagonal. Different lattices represent deformation to the regular square lattice as shown in Figure 2.9. Three scalars v1, v2, and v3, or one vector $\mathbf{v}$ (as shown in Figure 2.9), are used to define the lattice. For example, a square lattice has $\mathbf{v} = (1,0,1)$. Bitmap is a square lattice with no gap between pixels.

- Angular Resolution: $[0.1° - 2.7°]$
  For two light sources to be distinguishable, they must be separated by a certain angle. If the angle between them goes below a certain threshold, the rays from the sources overlap, and the sources become no longer distinguishable as shown in

Figure 2.10. For each light receiving hole, there is a characteristic limiting angle below which the light sources start to overlap. This angle is called the Angular Resolution. The Angular Resolution (also called Spatial Resolution) of the human eye, can be quantified by the Visual Angle between the two closest points that can be seen apart. The angular resolution of a healthy human eye is around 0.02°.

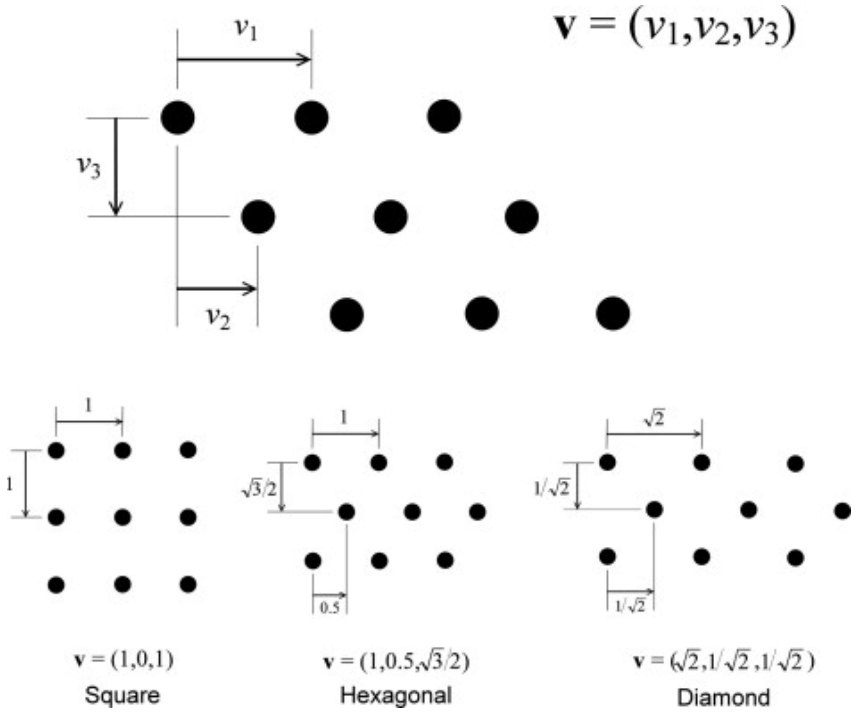


Figure 2.9: The shape of the visual field lattice [5].

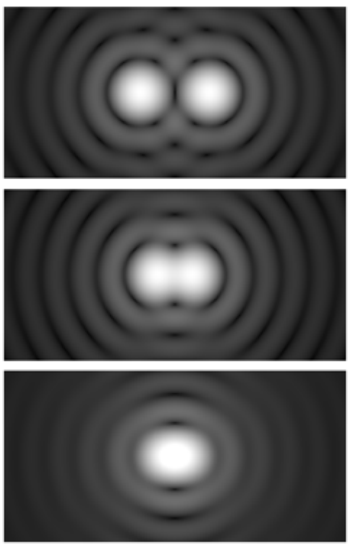


Figure 2.10: Closer light sources are less distinguishable [8].

- Field of view: $[1.7° \times 1.7° - 55° \times 41°]$ OR $[68px \times 68px - 2200px \times 1640px]$. (see equation 2.1 for visual angle to pixels conversion details)

- Number of grays: Most SPVs implemented fixed size phosphenes with variable luminance or variable size phosphenes with fixed luminance. These attributes had been modulated to levels between 2 and 32.

- Refresh rate: [10 Hz - 100 Hz]
  Refresh rate (in hertz): the number of times a monitor changes the displayed picture per second.
  Frame rate (in FPS): the number of images generated by a computer per second.

To draw a phosphene, image processing needs to be applied to the phosphene receptive field or aperture. The **receptive field** (or **phosphene aperture**) is the area on the image corresponding to the phosphene location on the phosphene map. A primitive image processing technique is **impulse sampling**: it is sampling gray-scale values at the receptive field to obtain the values for the phosphene PMIs. This can be achieved with a mean filter, or a Gaussian filter (preferential, weighted, or mean filter) applied on the receptive field [5].

2.3 Facial detection

Since this thesis aims (see section 1.1) at improving the ability of visual prostheses implantees in, not only recognizing familiar faces but also, seeing the facial details (e.g. nose, mouth, lips, etc.) clearly with the offered low resolution vision, facial detection in computer vision comes at the top of the most important concepts for this thesis. Past studies (see section 2.6), generally, worked on detecting faces in the high resolution real-time scene, and, with some processing applied on the detected faces, they managed to improve the ability of the subjects to recognize faces and see facial details. Therefore, this section presents an overview of one of the most notable face detection algorithms.

The face detection algorithm presented in this section is the **Viola Jones face detection algorithm**. This algorithm was used in most of the past studies, and is used in this thesis as well, to detect faces in the visual field. The algorithm works as follows: [9]

1. Create an image pyramid (i.e. multi-scale representation of the image as shown in Figure 2.11) to make the face detection process scale-invariant (i.e. to be able to detect large and small faces).

2. Create integral images: the value at a pixel location is the sum of the pixels that are above and left to that pixel. It helps in computing the sum of the pixels within a rectangular window in constant time as shown in Figure 2.12

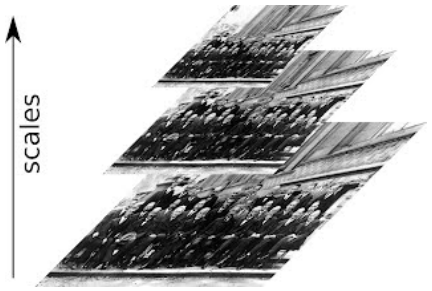


Figure 2.11: Image pyramid [9].

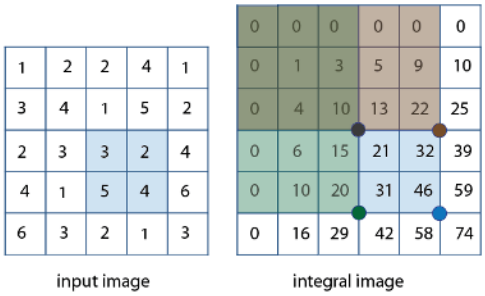


Figure 2.12: The summation of the shaded region pixels in the input image is an $O(1)$ computation using four reference values of the corresponding integral image. In the above example, $region\,pixels\,sum = 46 - 22 - 20 + 10 = 14$ [9].

3. Compute rectangle (Haar-like) features of different areas of the image using a sliding window. First, the pixel-sums in certain rectangles within the window are computed; this can be done in constant time with the help of the integral images. Then, subtracting some of these values can help in face detection. An example of rectangular features is depicted in Figure 2.13.

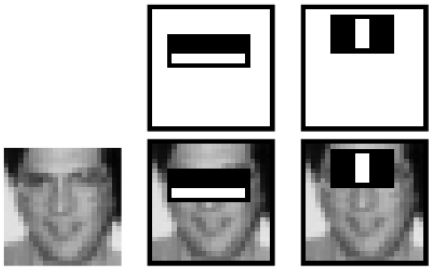


Figure 2.13: Examples of the rectangular features [9]. In the middle image, the depicted rectangular feature of the human face is that there is a darker rectangle around the eyes with respect to the rectangle below it (we can subtract these rectangles pixel-sums, and the result should be a negative value if the window contains a face). In the rightmost image, the rectangular feature is that there is a brighter rectangle around the nose bridge than the two rectangles around the eyes.

4. Scan the image pyramid of integral images with a square window (24×24 pixels) searching for areas that meet the rectangular features of the human face. To decide the rectangular features of the human face, a machine learning model was used. Around 160,000 features are present in a 24×24 pixels window, and only few of them are important in detecting faces. The ML model is trained using AdaBoost (i.e. making each feature act as a single classifier on the training data). After training, the model weights the features by their performance in detecting faces. Also, features that have performance below a certain threshold are rendered useless and are dropped by the model.

5. Through a series of weak classifiers (of features that showed the best performance in the training phase), as shown in Figure 2.14, determine whether the window contains a face or not. At each step, some rectangular features (with their weights taken into account) of the face are computed, and if the window passes, it moves on to further processing. Otherwise, the window is discarded. A window that passes all classifiers is decided to contain a face.

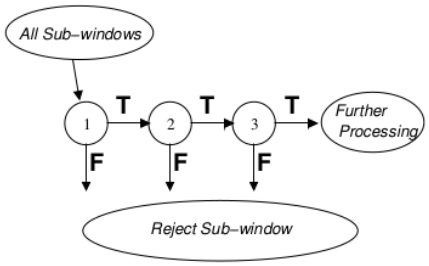


Figure 2.14: The cascade classifier of Viola Jones [9].

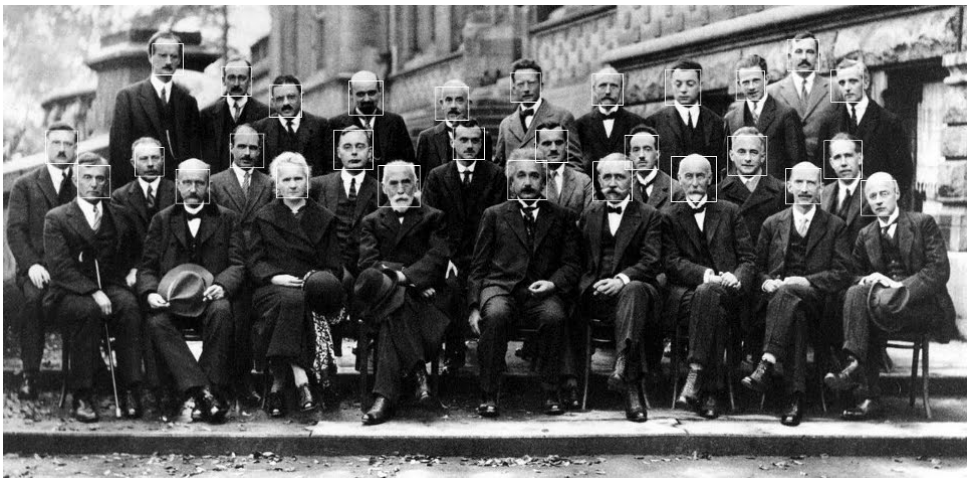


Figure 2.15: The Viola Jones algorithm detected face regions.

2.4 Face Landmark Detection

After detecting the regions that contain faces, further processing needs to be applied on the face regions. One of the most important concepts that help in processing a face region is face landmark detection. Face landmark detection (or face alignment) is a computer vision (CV) task where key-points from the human face are detected. Given the location and size of a face, the detected landmarks determine the shape of the face components such as eyes and nose [10]. This is particularly useful for this thesis in multiple face processing techniques like determining whether a person is talking or not based on the relative positions of their lips (see subsection 3.4.3), or determining the emotion that is most likely expressed by a person (see subsection 3.4.2). Face landmarks are also very useful in caricaturing faces which is explored in detail in section 2.5.

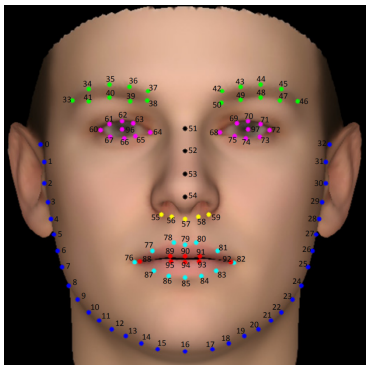


Figure 2.16: Face Landmark Detection [10].

2.5 Face Caricaturing

Caricaturing is one way to process a face region in order to make it more distinguishable. Caricaturing faces to help visual prostheses implantees was explored in the literature as outlined in section 2.6, and its implementation and usage within this thesis are outlined in subsection 3.4.4. In this section, a theoretical overview of caricaturing is presented.

As described in the literature [11], one way faces can be coded is via a perceptual *face-space*, where faces are coded in terms of how their particular characteristics differ from the average face as shown in Figure 2.17. The average face lies at the center of the *face-space*, and each individual face lies in a location determined by its values on multiple dimensions that vary between real-world faces. The density of faces is lower in peripheral than more central regions of the space because values on dimensions are normally distributed in the real-world faces that provide the input on which perceptual face-space is formed. *Face-space* coding implies that the ability to tell apart individual faces can be improved by exaggerating the way a veridical face differs from the average face. This procedure is called Caricaturing, and is illustrated in Figure 2.17.

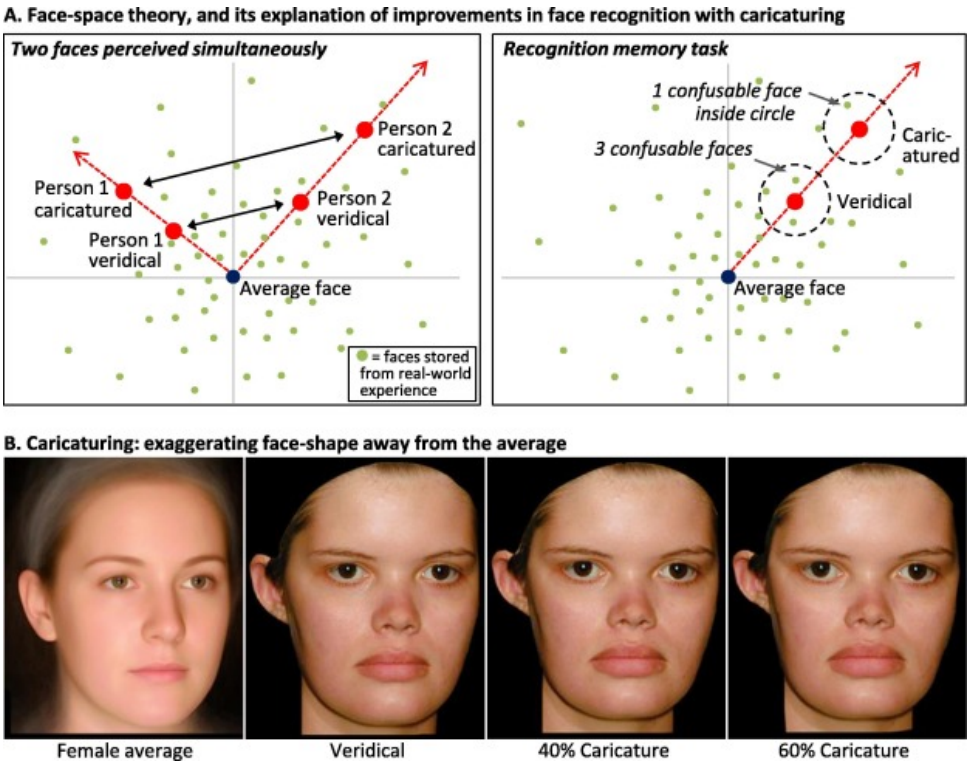


Figure 2.17: **A.** Perceptual face-space (i.e., how faces are represented in the brain) and how it explains improved face recognition with caricaturing. The dimensions coded on the axes remain unknown but might represent, for example, lip thickness or face width. **B.** Example of face caricaturing: altering the veridical face physically away from the average. Average face is made by averaging together faces of multiple people [11].

Face Caricaturing can improve perception of the differences between pairs of simultaneously seen faces, and it can be applied on the faces using the opposite of Face Morphing techniques. Face Morphing can be done in the following way: given two input images I_0 and I_1 of human faces, morphing is generating a fluid transformation (video clip) transitioning from I_0 to I_1 as shown in Figure 2.18. Applying the opposite of Face morphing (i.e. transitioning a veridical face *away* from an average face) is one way of doing caricaturing.

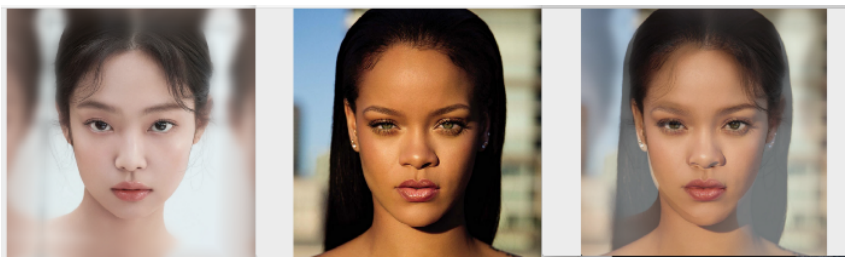


Figure 2.18: An example of Face Morphing. Left: input image I_0. Middle: input image I_1. Right: cross-dissolve of I_0 and I_1.

2.6 Related Work

Several past studies explored different processing strategies to facial regions and the impact of these strategies on improving the visual prostheses implantees (or SPVs subjects) ability to recognize faces. The explored strategies were increasing contrast [12], region-of-interest (ROI) magnification [13], and caricaturing [11].

2.6.1 Histogram Equalization

Increasing image contrast is one of the most common image enhancement techniques. According to this study [12], it helped in increasing the ability of the SPVs subjects in recognizing faces. Increasing image contrast can be achieved through histogram equalization [14]. An **image histogram** is a graphical representation of the intensity distribution of an image. Histogram equalization increases image contrast by stretching out intensity range. An example of histogram equalization is shown in Figure 2.19.

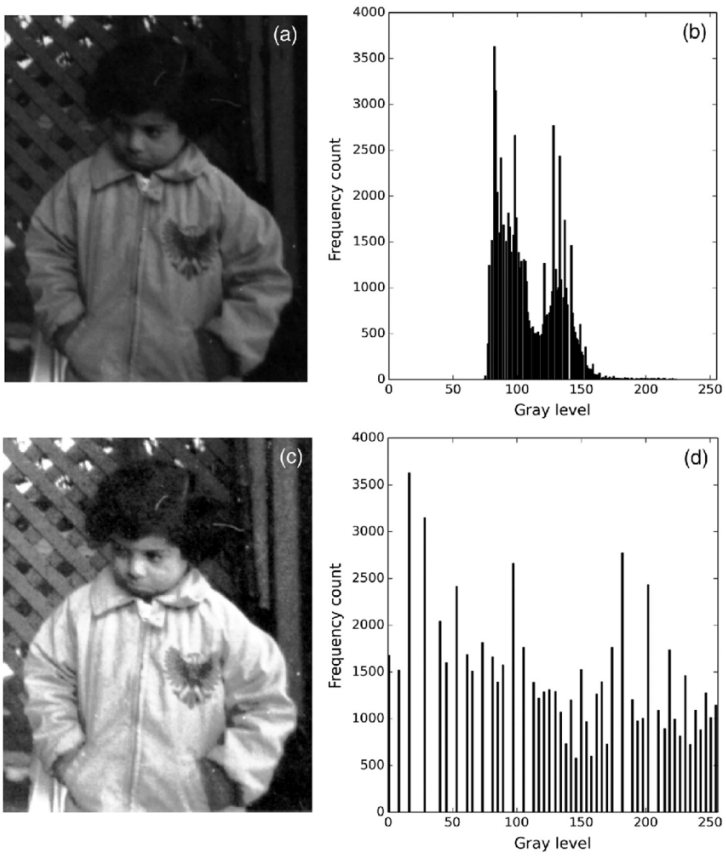


Figure 2.19: Top row: input image with its image histogram. Bottom row: equalized image with its image histogram.

2.6.2 Region of interest magnification

Because the visual field of the visual prostheses implantees has only a very limited number of phosphenes, these phosphenes can only carry limited information. One way to make the phosphenes carry as much useful information as possible is region-of-interest (ROI) magnification. The ROI could be the Viola Jones face region or any other region.

According to this paper [13], the ROI magnification shows promising results in increasing the SPVs subjects face recognition ability. The ROIs explored in the study were the following:

- Viola Jones face region (VJFR) [9]
- Statistical face region (SFR) is obtained through:
    1. Nose position detection (NPD) that is computed through: Otsu's thresholding and connected component analysis (to find distinct components in an image) on VJFR. The nose is a distinct component on the face, and the nose position can serve as a centering landmark to enlarge VJFR with proper proportions.
    2. Enlarging the VJFR window according to statistical enlarging ratios. These ratios are precalculated and are applied with respect to the position of the nose to include more features in the region like hair and ears.
- Matting face region (MFR) is obtained through matting. Matting is the extraction of foreground from background. This is basically useful when the SFR contains background unneeded information that can be discarded.

The applied SPV in the study had the following attributes:

- Gaussian phosphenes
- 8 gray levels
- 24×24 OR 32×32 phosphenes with dot patch (dot and spacing) size $0.75°$ or $0.56°$ respectively

A summary of this past study processes is shown in Figure 2.20. Although MFR-ROI magnification achieved the highest recognition accuracy by SPV subjects, it had the highest algorithmic complexity and cannot fulfill real-time processing requirements. The study, therefore, suggests that SFR-ROI magnification is the most favorable technique taking into consideration two parameters: subjects recognition accuracy, and processing speed.

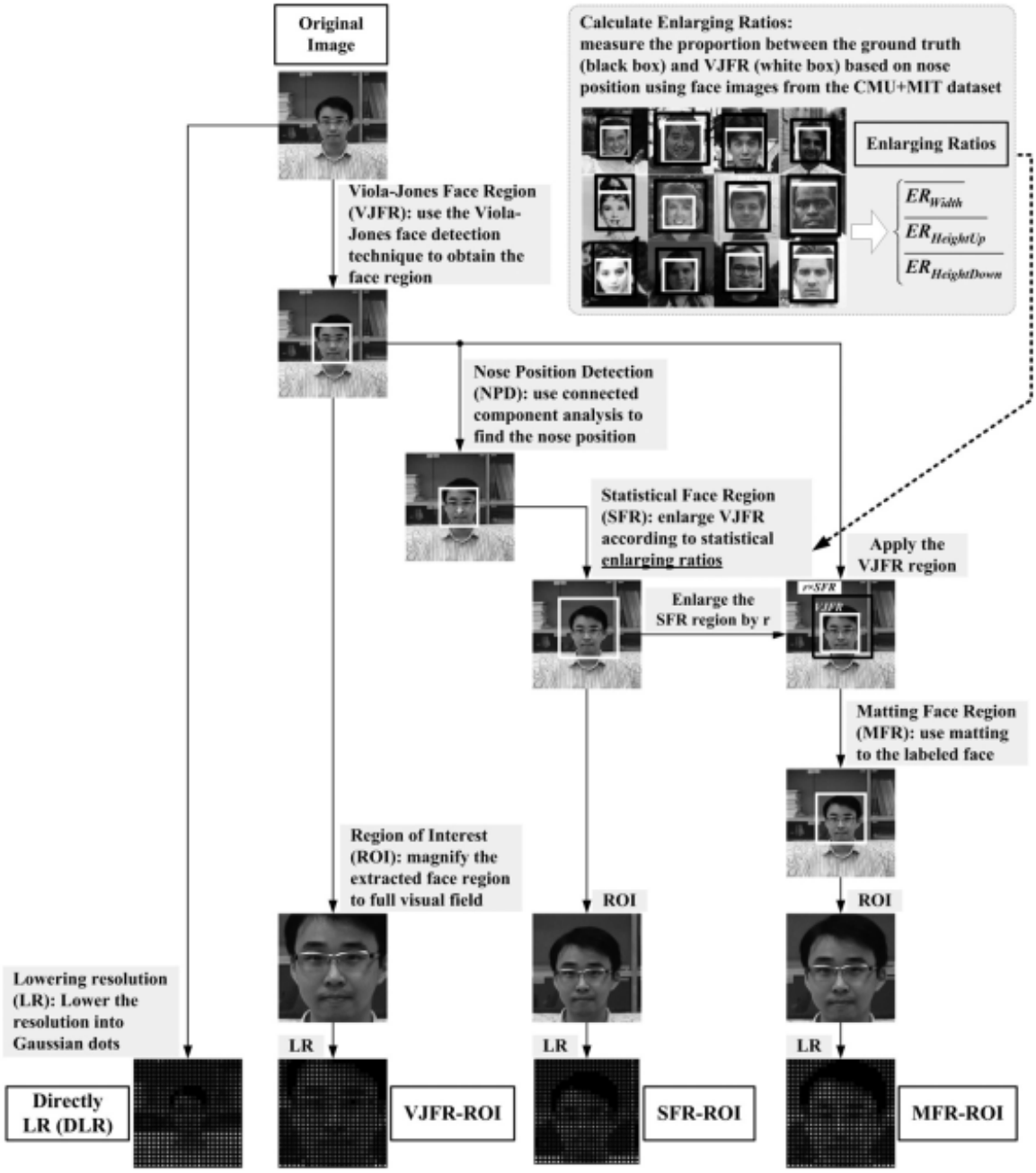


Figure 2.20: ROIs magnification in SPVs [13].

2.6.3 Caricaturing

A past study [11] suggests that its previous studies had only reported high face recognition accuracy by relying heavily on details that survive low resolution well. These details include general category information (such as sex, race, or age), and extra-face information (glasses, facial hair, and hairstyle). It is also noted that removing hair information from images, despite the fact that this increased the resolution of face region

(the same number of phosphenes will be portraying only the face region, rather than the whole head), reduced face recognition accuracy. Therefore, within-category discrimination of individuals has to be explored in prosthetic vision where only face information is informative, such as telling apart several young adult Caucasian men with similar hairstyles.

Although face category and extra-face information can be important in recognizing faces in everyday life, they are unreliable in many real world settings. For example, they would be of little help at a business meeting where all attendees are middle-aged men with short dark hair and are all dressed similarly. In addition, extra-face cues are constantly changing: a man might shave his mustache, or a woman might cut her long hair. Finally, these cues are useless when the task is to recognize a person from a large pool of possibilities. For example, telling whether a person approaching you in the street is someone you know.

Inspired by cochlear implant signal manipulations that were designed specifically to match the way the humans perceive speech, manipulations designed specifically to enhance humans perception and recognition of faces is explored. One possible manipulation is face caricaturing. A theoretical overview of face caricaturing is presented in section 2.5. This past study results suggest that caricaturing may be a useful method to enhance visual prostheses implantees performance in recognizing faces.

Caricaturing faces imposes so many challenges, as it cannot meet real-time requirements without greatly jeopardizing quality. A high-quality caricaturing algorithm shall first decide what average face to caricature away from rather than having a single average face, as this is the case in the implementation of this thesis (see subsection 3.4.4). The average face selection must be based on the veridical face age, sex, race, viewpoint, and expression. Also, an accurate caricaturing algorithm would need more than 68 facial landmarks, as this is the case in the implementation of this thesis, to work on. A possible better approach for caricaturing faces that can be investigated in future studies is making use of generative adversarial networks (GANs), like CariGAN [15].

Chapter 3

Methodologies

In order to compare the efficiency of different face region processing strategies described in the literature (see section 2.6) in achieving the thesis objectives (see section 1.2), an SPV, applying these strategies, was implemented [16] and was experimented on normally sighted subjects on both a computer screen and an HMD. The simulation project is written in **Python** because it is easy to learn and use with simple syntax, and it also has a huge packages index called **PyPi**. The packages supported offer great tools with easy-to-use APIs for optimization (e.g. **Numpy**), for Computer Vision (e.g. **OpenCV**), and for Machine Learning (e.g. **Dlib**). In this chapter, the implemented simulation and the conducted experiments setup are described in detail. The implemented simulation flow chart is presented in Figure 3.1.

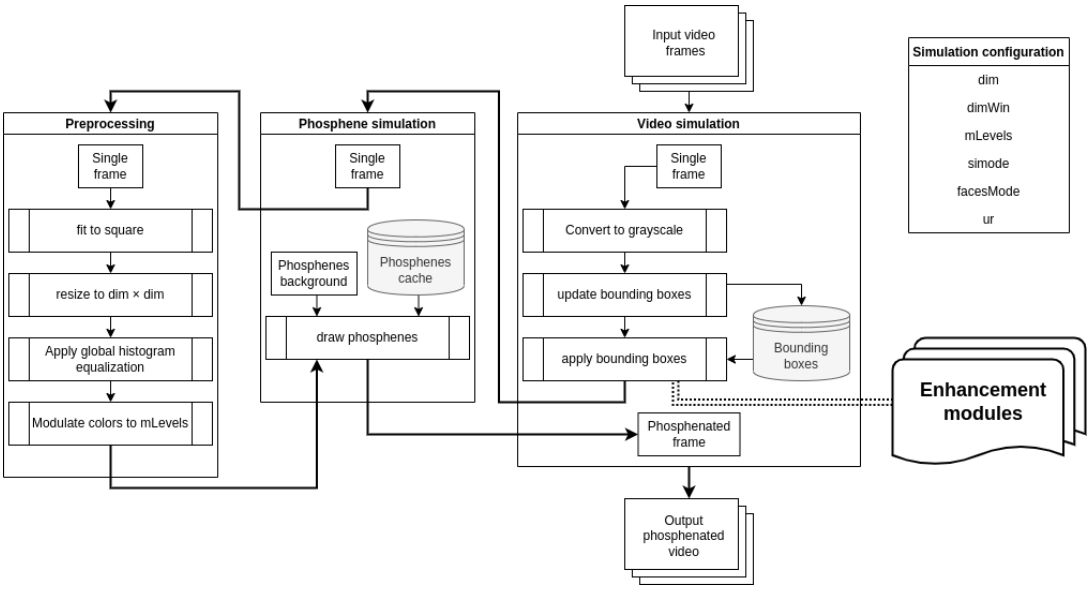


Figure 3.1: Simulation flowchart

3.1 Simulation Configuration

The SPV has some configuration variables that control the behavior of the simulation and can be set by the SPV user. These configuration variables are listed below:

1. **dim** *(32 by default)*: the number of phosphenes per one dimension. In other words, output phosphenated images will consist of **dim** × **dim** phosphenes.
2. **dimWin** *(640 by default)*: the size of the output images in pixels. In other words, generated images will have a resolution of **dimWin** × **dimWin** pixels.
3. **mLevels** *(16 by default)*: the number of modulation levels (the number of available colors or sizes) of phosphenes.
4. **simode**: the simulation mode. It is described in detail in subsection 3.5.1.
5. **facesMode**: the face enhancement mode. See subsection 3.5.2 for details.
6. **ur**: the update rate of the simulation enhancement. Because most of the enhancement options are computationally expensive, they are not applied on every frame but are rather applied once every **ur** frames.

The SPV has some default attributes that are not to be altered:

- Phosphene color: gray-scale
- Phosphene shape: Gaussian Distribution
- Lattice shape: square (see Figure 2.9)

3.2 Video Simulation

As shown in Figure 3.1, the SPV accepts as an input a video clip. It can also accept video feed from the web-cam to run in a real-time setting. Videos are phosphenated (i.e. transformed from pixels to phosphenes) one frame at a time in the SPV. Before a frame is mapped to the corresponding phosphene map, enhancements are applied to that frame. These enhancements are applied in two consecutive steps described in detail in the following subsections. Before applying any enhancements, the frame is converted to gray-scale. The frame three color channels are useless because the Viola Jones face detection algorithm works on gray-scale images [9] and because the phosphenes are all in gray-scale, so there is no need to proceed with a colored frame (i.e. proceeding with a gray-scale frame will have no negative impact on the subsequent computations but will rather make them simpler by lowering the space complexity).

3.2.1 Update Bounding Boxes

Since all enhancement methods apply Viola Jones Face detection then proceed with further processing on face regions, the face detection step could be separated from the rest of the SPV logic. **Updating bounding boxes** is basically letting the *Viola Jones Haar Cascade Classifier* detect faces in a passed frame. The bounding boxes of the detected faces are then stored (as shown in Figure 3.1) and are only updated once every **ur** frames (see section 3.1 for details). Figure 3.2 shows what it would look like if the *Haar Cascade Classifier* computed bounding box is drawn on the image passed to it.

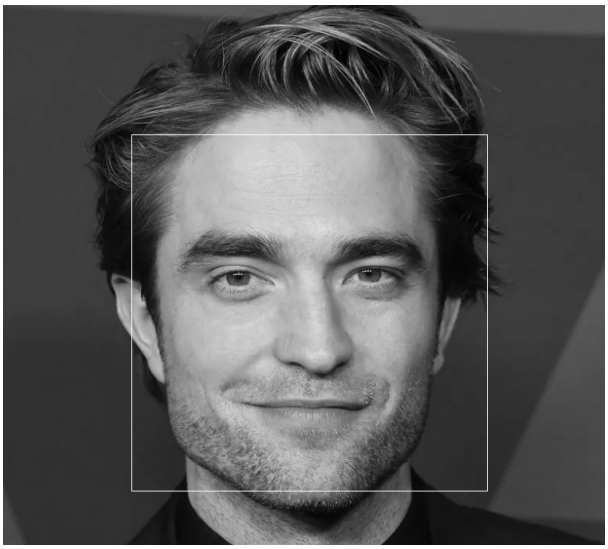


Figure 3.2: The Haar cascade classifier generates bounding boxes around detected faces.

Updating the bounding boxes once every **ur** frames does not only lower the computational cost of the face detection, but it also provides a more stable visual field for the subjects since all enhancement methods include magnification of the face region (see subsection 3.5.2 for details). For example, imagine a scenario where the displayed face is nodding. Updating the bounding boxes once every frame will result in a very unstable visual field, as the visual field will keep tracking the moving face so frequently. A more idle update rate would offer more stability to the visual field for the subjects.

3.2.2 Apply Bounding Boxes

In this step, the stored bounding boxes are fetched and are used to apply the configured enhancements to the face regions (as shown in Figure 3.1). This is applied on every frame of the video clip, and not once every **ur** frame like the previous step. The applied enhancements to the face regions are dictated by the configuration variable **facesMode** (see section 3.5.2 for details). In addition, the face regions could be transformed in this step from being VJFR to the SFR mentioned in the literature (see subsection 2.6.2) also based on the SPV configuration.

An imposed challenge in this step was having more than a single bounding box in a frame (i.e. having more than one face visible in the visual field of the subjects), which is not an unlikely setting. Since one face has to be chosen to be enhanced then to fill in the visual field, the subjects must have the flexibility to switch between visible faces. A global variable can be used to represent the index of the face of interest, and with a controller, the subjects can switch between faces in a cyclic fashion. Bounding boxes indices example is shown in Figure 3.3.

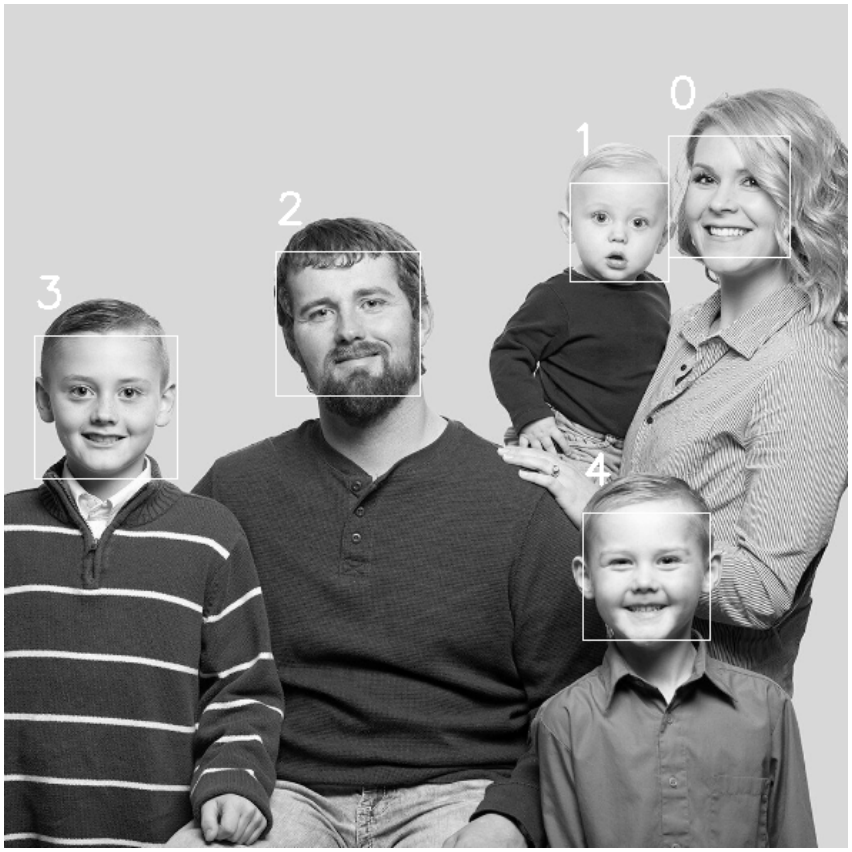


Figure 3.3: The Haar cascade classifier generates bounding boxes around detected faces. In the above example, a global variable $index_{face}$ that references the face of interest is initially set to 0, and the user can change its value up to 4 then back again to 0. The referenced face is the one that is enhanced then used to fill the visual field.

3.3 Singleton Image Simulation

After the bounding box of the face of interest is detected and enhanced then magnified, the frame is now phosphenated. A blank image (phosphenes background shown in Figure 3.1) is created, and is used as a blank canvas to draw phosphenes. Before mapping pixels to phosphenes, the frame undergoes preprocessing.

3.3.1 Image Preprocessing

The first step in preprocessing is to change the frame dimensions to make them equal (i.e. convert a rectangular image into a square one). This is necessary because the SPV was designed to generate square phosphenated images. This can be done in one of two ways: **Square Crop** and **Square fit** as shown in Figure 3.4. **Square fit** approach was adopted in this thesis because the **Square Crop** approach discards information from the visual field, which is an undesired behavior.

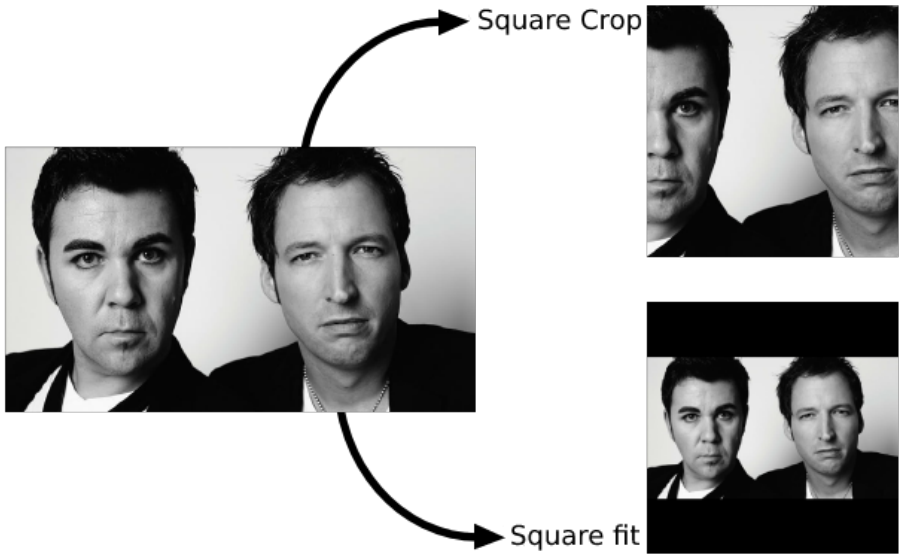


Figure 3.4:
Square crop: crops the image longer dimension to be equal to the shorter one.
Square fit: fill with black pixels the shorter dimension to be equal to the longer one.

In order to compute the gray level, or the size in case of size modulated phosphenes, of a single phosphene, two approaches were explored. Both the approaches included dividing the image into patches, or **receptive fields**, and then followed by operating on these fields. The first approach was to take a sample in the middle of the patch to be assigned as the gray level of the corresponding phosphene. This approach, similar to **nearest neighbor interpolation** approaches, despite having the lower computational complexity, disregarded a lot of information from the input image. The second approach, and the approach adopted in the second step of preprocessing, is resizing the image using the **bilinear interpolation** approach. The output image of this step contains as many pixels as there should be phosphenes (i.e. the resized image will be $dim \times \text{dim } px$ OR $32 \times 32\,px$ by default) (see section 3.1 for details). Applying **bilinear interpolation** in resizing the image is very similar to **Impulse Sampling** mentioned in the literature, as it is basically a weighted filter. This does not disregard any information from the input image, unlike what the **nearest neighbor interpolation** does.

The next step in preprocessing is increasing frame contrast using **Histogram Equalization**. As mentioned in this subsection 2.6.1, histogram equalization is an image enhancement technique that increases image contrast. This is useful in overcoming any unclarities in the frame caused by poor lighting conditions.

The next, and the last step in the preprocessing of the frame, is color modulation. In this step, the intensities of the pixels in the frame are modulated changing their range from 256 different values to only **mLevels**, or 16 by default (see section 3.1 for details), different values. This is done in two steps:

1. The color of each pixel is transformed from the range $[0 - 255]$ to the new range $[0 - (mLevels - 1)]$, or $[0 - 15]$ by default, according to the following equation of **Linear Normalization**:

$$tmp = integer(input \times \frac{mLevels - 1}{255}) \tag{3.1}$$

   where *input* is the input pixel intensity, and *integer*() is a function that rounds to the nearest integer, and *tmp* is an intermediate result.

2. The intermediate result is then transformed back, through **Linear Normalization** as well, to the original domain of $[0 - 255]$ through this equation:

$$output = integer(tmp \times \frac{255}{mLevels - 1}) \tag{3.2}$$

This way, each pixel will contain a value between 0 and 255, and the number of distinct pixel values in the frame will be equal to *mLevels*.

3.3.2 Drawing Phosphenes

In this step, the preprocessed frame is used to draw phosphenes on the phosphenes background (see Figure 3.1). All the pixels of the phosphenes background are initially set to zero (i.e. the background is initially a black image). The background is divided into a number of squares equal to the number of phosphenes (equal to the number of pixels in the preprocessed frame). The pixels of the preprocessed frame are traversed pixel by pixel, and using their modulated intensity values, the SPV draws corresponding phosphenes on the corresponding squares on the background as shown in Figure 3.5. For a pixel located at coordinates (x_i, y_i) on the input image, the coordinates of the top left corner of its corresponding square on the phosphenes background (X_i, Y_i) are computed as follows:

$$square_{side} = \frac{dimWin}{dim}$$

$$(X_i, Y_i) = (x_i \times square_{side}, y_i \times square_{side}) \tag{3.3}$$

where $square_{side}$ is the length of square side that will contain a single phosphene, and $dimWin$ and dim are SPV configuration variables (see section 3.1 for details).

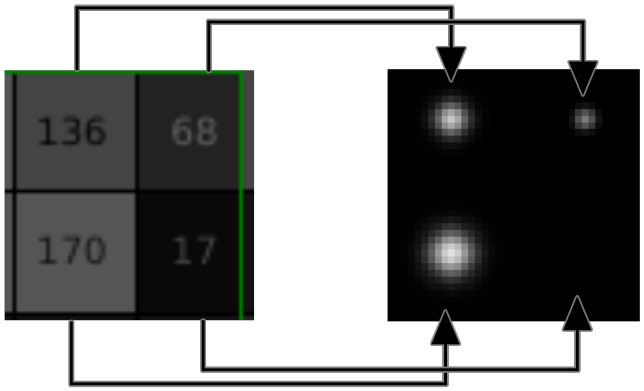


Figure 3.5: The conversion of pixels to size modulated phosphenes.

Knowing the pixel color and the corresponding top left corner coordinates of the square that will be used to draw a phosphene, the phosphene can be drawn in one of four ways:

1. Blur Color Modulated (*BCM*)
2. Blur Size Modulated (*BSM*)
3. Array Color Modulated (*ACM*)
4. Array Size Modulated (*ASM*)

There is a dedicated subsection (see subsection 3.5.1 for details) to explain these four drawing methods in detail. In addition, once a phosphene is drawn, it is cached because the phosphene drawing methods are computationally complex. When the SPV tries to draw the corresponding phosphene of a certain color of a pixel, the cache is checked initially for matches. If the needed phosphene was found, it is directly fetched from the cache. If the phosphene was not found, it is drawn then cached. An example of a phosphene cache is shown in Figure 3.6. After all pixels are used to draw corresponding phosphenes, the output image is now ready to be displayed to the subjects.

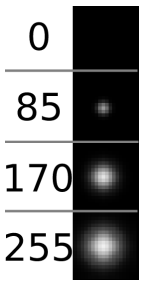


Figure 3.6: An example of a phosphene cache with four size modulation levels. The phosphenes in the cache are indexed by the color that was used to draw them.

3.4 Enhancement Modules

Multiple face enhancement modules (each module implementing a face enhancement strategy) are integrated in the SPV. Most of these modules share two necessary steps:

1. The bounding boxes of the faces are detected using the *Viola Jones Haar Cascade Classifier* (see section 2.3 for details).
    - **Implementation**: the **Python OpenCV** library offers a pretrained *Haar Cascade Classifier* that can be used with a very simple **API**.
2. Within each face bounding box, **Face Landmark Detection** (see section 2.4 for details) is applied. Sixty-eight face landmarks are detected using the *One Millisecond Face Alignment with an Ensemble of Regression Trees* algorithm [10] as shown in the Figure 3.7.
    - **Implementation**: the **Python dlib** library offers a pretrained *Face Landmarks Predictor* that can be used with a very simple **API**.

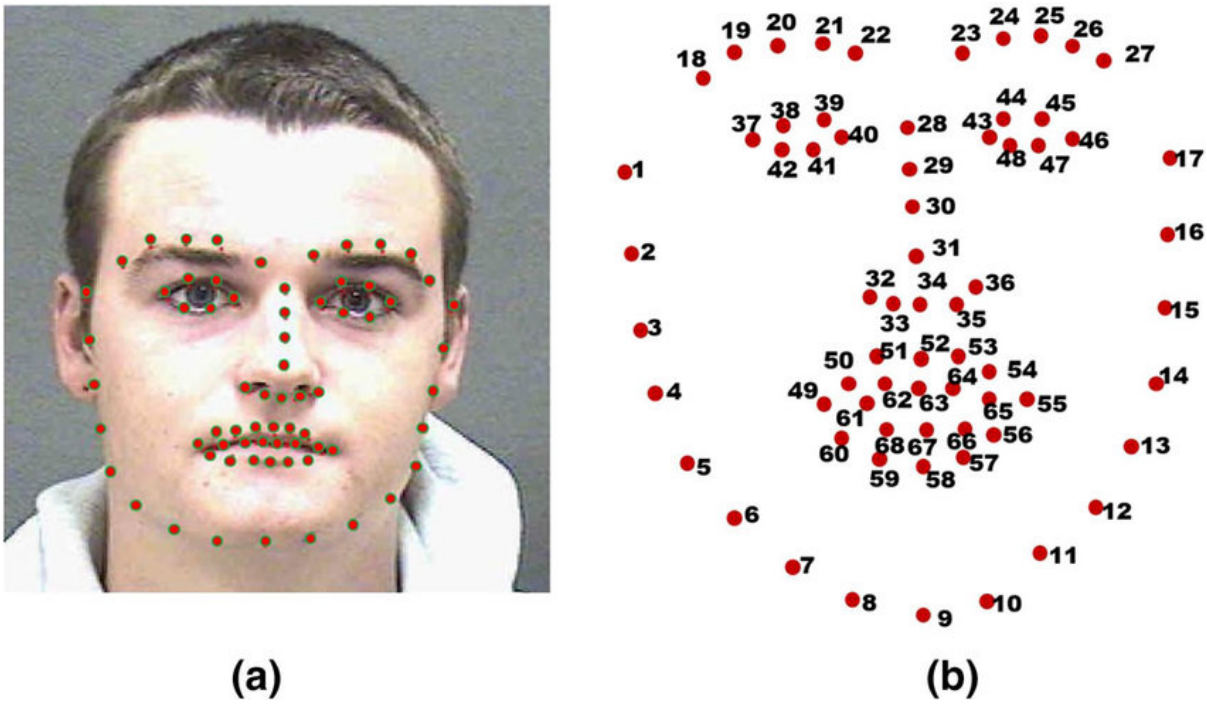


Figure 3.7: Identification of 68 facial landmarks using Dlib.

After assigning the face of interest (see subsection 3.2.2 for details), each module proceeds with further processing on the *Face Bounding Box* and the *Face Landmarks*.

3.4.1 Face-specific histogram equalization (FSHE)

This module takes as an input a single face bounding box with its landmarks and applies histogram equalization to only the face area. To apply face-specific histogram equalization, first the minimum polygon containing all the 68 **Dlib** landmarks is computed. Histogram equalization is then applied to that polygon, in addition to the histogram equalization applied to the whole image (see subsection 3.3.1 for details), in order to overcome any unclarities in the facial details caused by poor lighting conditions. Figure 3.8 shows an example of face specific histogram equalization.

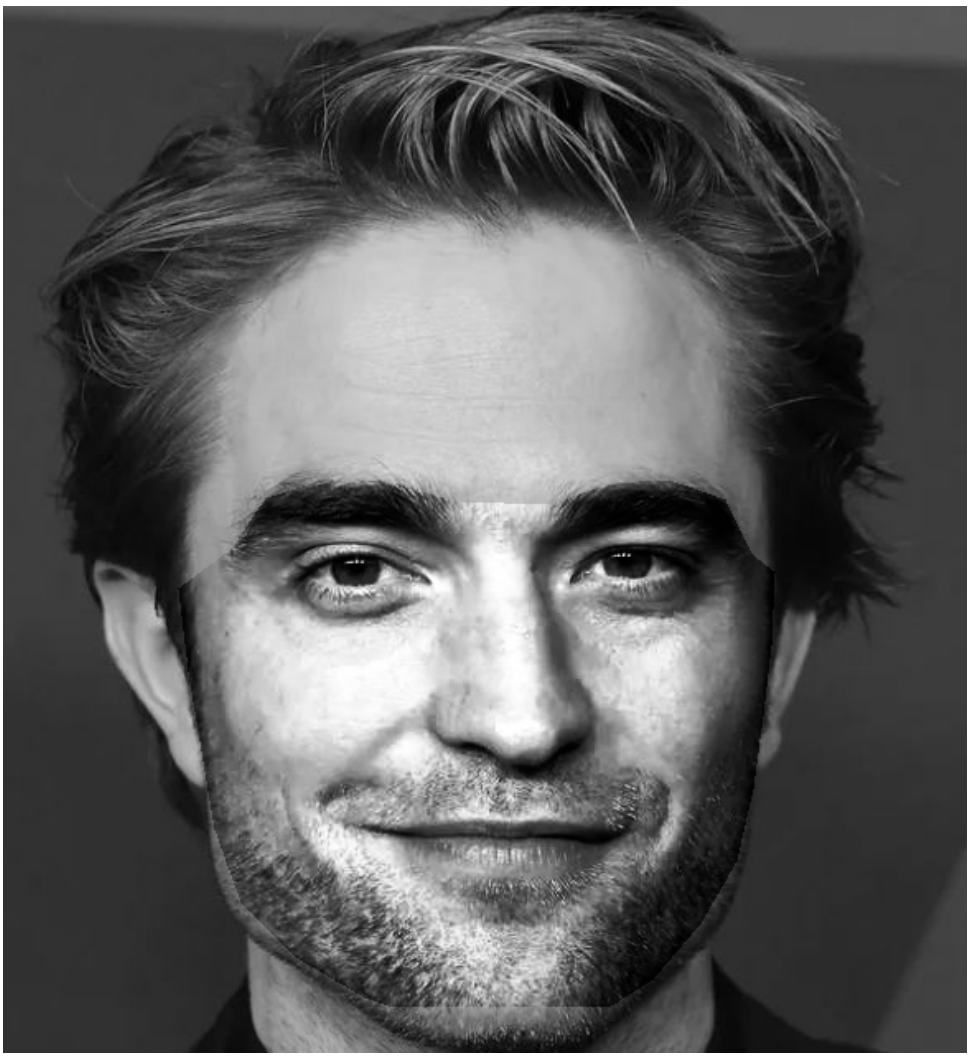


Figure 3.8: Face Specific histogram equalization (FSHE) is HE applied on the minimum polygon containing the **Dlib** 68 face landmarks.

3.4.2 Emotion Recognition (ER)

Taking as an input the *Viola Jones Face Region*, this module, making use of a neural network, outputs the emotion that is most likely expressed by the passed face. The possible emotions that this module can recognize are the following:

1. angry,
2. disgusted,
3. fearful,
4. happy,
5. neutral,
6. sad,
7. and surprised.

A pretrained deep convolutional neural network (CNN) is used to implement this module. The **GitHub** repository on which this CNN was found is in this url: https://github.com/atulapra/Emotion-detection.

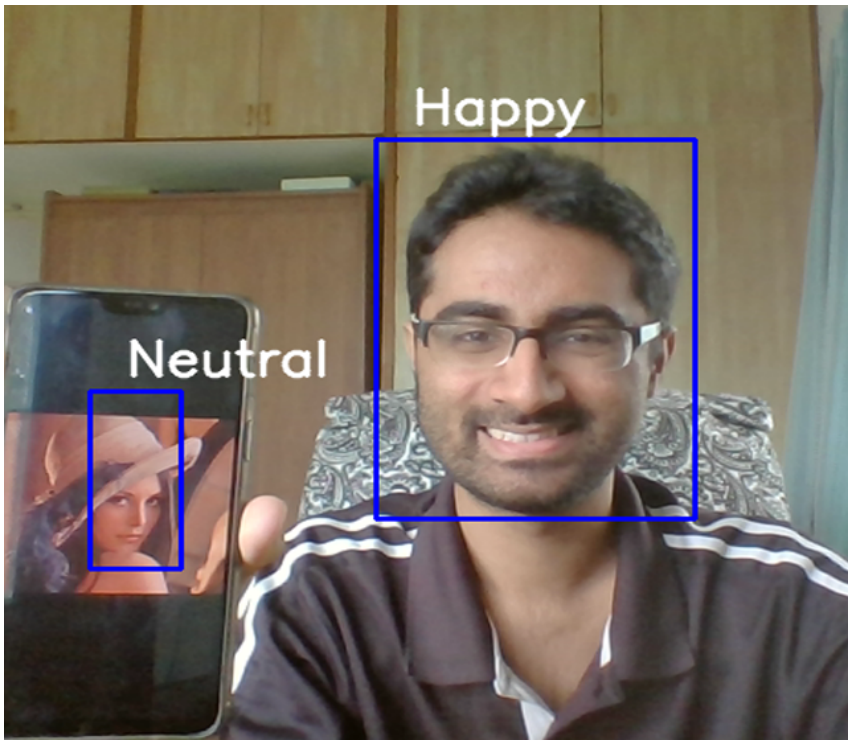


Figure 3.9: Sample output of the Emotion Recognition module.

3.4.3 Talking Detection (TD)

This is a recurrent neural network (RNN) based module that takes as an input an array of 25 consecutive frames of the same *Viola Jones* face. On each frame, **Dlib Face Landmark Detection** is applied. A $mouth_{gap}$ distance from these landmarks is then calculated according to the following equations:

$$A = dist(point_{61}, point_{67})$$

$$B = dist(point_{62}, point_{66})$$

$$C = dist(point_{63}, point_{65})$$

$$mouth_{gap} = \frac{A + B + C}{3} \tag{3.4}$$

where $dist(point_i, point_j)$ is a function that returns the *Euclidean Distance* between the two points, $point_i$ and $point_j$, and each point referenced in the equations represents one of the points of the **Dlib** 68 face landmarks (e.g. $point_i$ represents the i^{th} landmark). Depiction of the $mouth_{gap}$ calculation is shown on Figure 3.10.

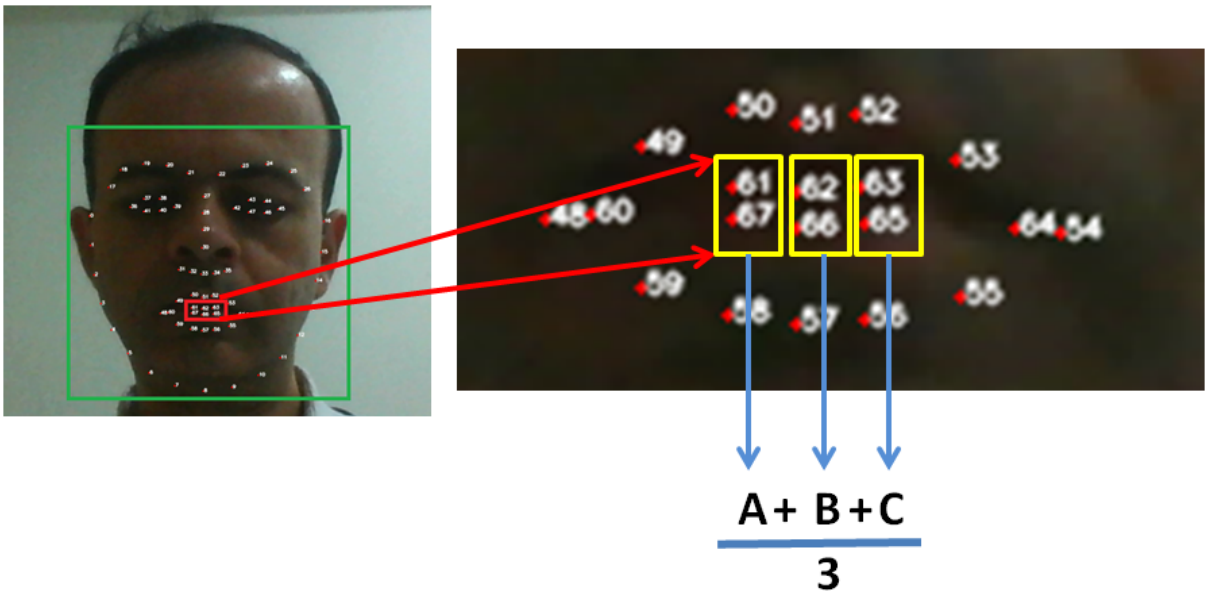


Figure 3.10: Calculating the $mouth_{gap}$ for the talking detection RNN.

This will result in an array of 25 values for the 25 frames $[mouth_{gap_1}, ..., mouth_{gap_{25}}]$. The values in the array are then normalized (i.e. their range is modified to be $[0 - 1]$). The RNN accepts this normalized array of $mouth_{gap}$ values, and returns the probability of that person being speaking and the probability of them being silent. The pretrained recurrent neural network (RNN) was found on this **GitHub** repository: https://github.com/sachinsdate/lip-movement-net.

3.4.4 Caricaturing

As mentioned in section 2.5, caricaturing can be applied on a face region by doing the reverse of face morphing. A **GitHub** repository implementing face morphing was found, and this implementation was altered in this module because the goal was to do **Face Caricaturing** not **Face Morphing**. The repository source code accepts as an input two face images I_0 and I_1, and it outputs a fluid transformation video clip from I_0 to I_1. The **Caricaturing** module, in contrast, accepts as an input a single face image, and it outputs the result of transforming this face **away** from a stored average face within the module. An implementation of **Face Morphing** was found in this **GitHub** repository: https://github.com/Azmarie/Face-Morphing.

The module, making use of the source code found in the aforementioned repository, does the following:

1. At the SPV startup, it computes the Average face landmarks to use them as a reference for **Caricaturing** passed faces as shown in Figure 3.13 B.

2. For a veridical passed face, it computes its landmarks and converts them into a **Delaunay Triangulation** mesh as shown in Figure 3.13 D. **Delaunay Triangulation** for a given set P of discrete points in a general position, is a triangulation DT such that no point in P is inside the circumcircle of any triangle in DT as shown in Figure 3.11.

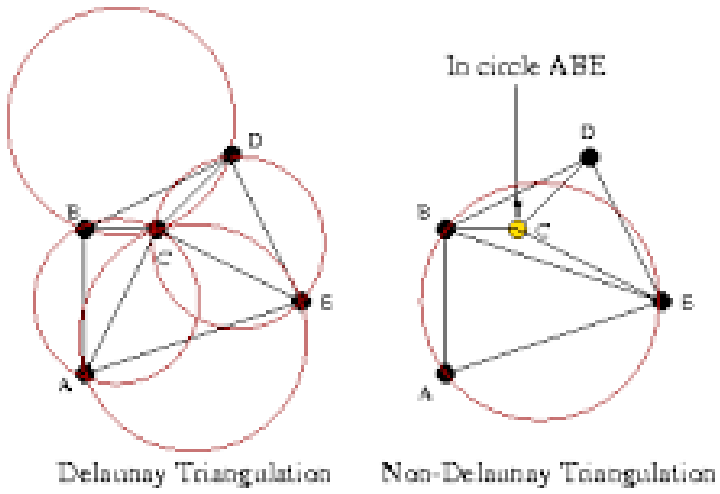


Figure 3.11: Delaunay vs Non-Delaunay triangulation [17].

3. For the i^{th} landmark on the veridical face (point lm_v^i with coordinates (x_v^i, y_v^i)), the module strays it away from its corresponding landmark on the average face (point lm_a^i with coordinates (x_a^i, y_a^i)) with a factor α, and the resulting point P^i is stored. According to the literature [11], an optimal value for α is 60%. Figure 3.12 shows

how lm_v^i is strayed away from lm_a^i. First, a new point lm_n^i with coordinates (x_n^i, y_n^i) is computed as follows:

$$x_n^i = 2 \times x_v^i - x_a^i$$
$$y_n^i = 2 \times y_v^i - y_a^i \tag{3.5}$$

Using lm_n^i, a new stray point P^i with coordinates (x_P^i, y_P^i) is computed as follows:

$$x_P^i = (1 - \alpha) \times x_v^i + \alpha \times x_n^i$$
$$y_P^i = (1 - \alpha) \times y_v^i + \alpha \times y_n^i \tag{3.6}$$

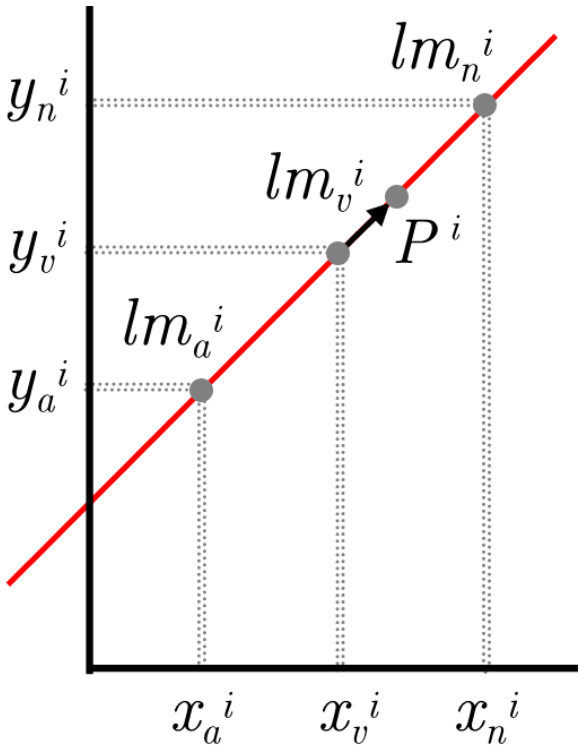
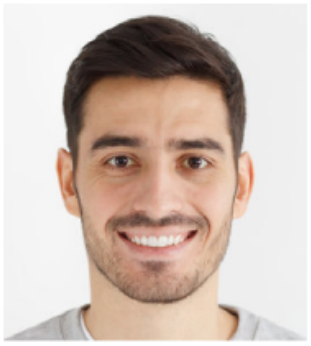


Figure 3.12: To stray lm_v^i away from lm_a^i, a new point lm_n^i is computed on the other side of the straight line joining the two points such that lm_v^i lies in the middle between lm_a^i and lm_n^i. By moving lm_v^i towards lm_n^i with a factor α, the point P^i is obtained for the i^{th} landmark of the face.

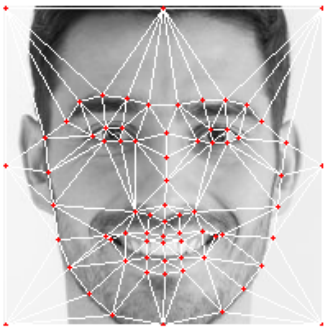
4. **Affine Transformations** (geometric transformations that preserve lines and parallelism) are applied on each triangle in the **Delaunay Triangulation** mesh transforming each triangle $[lm_v^i, lm_v^j, lm_v^k]$ into a new triangle $[P^i, P^j, P^k]$ (i.e. moving each veridical landmark lm_v^i towards its corresponding stray point P^i). The resulting collective image of transformed triangles is the output caricatured face image. A face caricaturing example is shown in Figure 3.13 E.

A. Average face

B. Stored Average face landmarks



C. Veridical face



D. Veridical face landmarks and Delaunay triangulation

E. Final Caricatured face

Figure 3.13: The steps of caricaturing a face with $\alpha = 60\%$.

3.5 Supported Modes

The simulation configuration variables (see section 3.1 for details) can be altered to control which simulation mode to use and which enhancement mode to apply. A list of these modes is explored in detail in this section.

3.5.1 Simulation Mode (*simode*)

The *simode* configuration variable tells the simulation two things: how to render different phosphene modulation levels, and how to draw a Gaussian phosphene. First, the phosphenes could be either **Size** modulated (with a fixed color), or **Color** (gray-scale) modulated (with a fixed size). Second, the simulation can draw a Gaussian phosphene in one of two ways:

1. *Blur* method. To draw a phosphene using this method:
   (a) Draw a solid circle.
   (b) Apply a Gaussian convolution kernel (as shown in Figure 3.14).

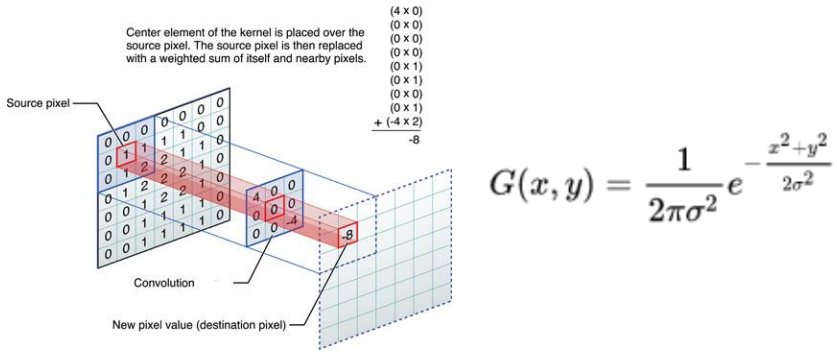


Figure 3.14: A Gaussian Convolution Kernel is a convolution kernel whose elements represent the shape of a Gaussian hump (i.e. calculated from the gaussian function in 2D). In image processing, a kernel is a small matrix, and convolution is adding each element of the image to its local neighbors weighted by the kernel [18].

2. *Array* of circles method. To draw a phosphene using this method:
   - Draw multiple circles around the phosphene center. Further circles from the center have a darker color. Circles colors are calculated by obtaining a factor β from the 1D Gaussian distribution as shown in Figure 3.15. Then the equation

$$circle^{i}_{color} = \beta^{i} \times 255 \tag{3.7}$$

is used to compute i^{th} circle color. Size modulated phosphenes have a varying number of constituting circles, and color modulated phosphenes have the same number of circles but different amplitudes (i.e. β^i is multiplied by different values and not 255 for every phosphene).

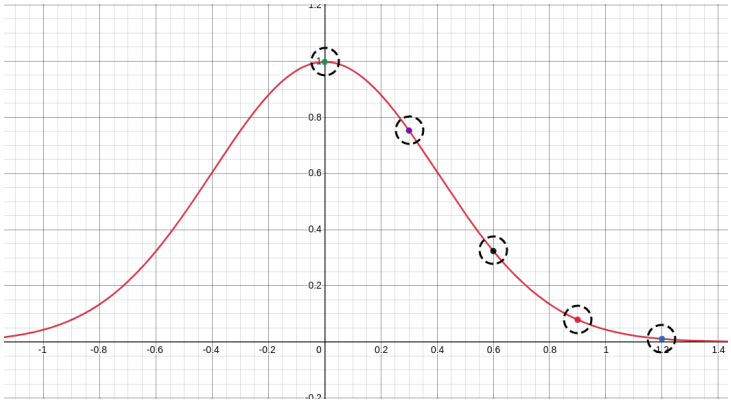


Figure 3.15: For a phosphene comprising five circles, five linearly equidistant β values are used from the Gaussian distribution to compute each circle color.

Accordingly, this configuration variable instructs the simulation to operate in one of four modes:

1. Blur Color Modulated (*BCM*)
2. Blur Size Modulated (*BSM*)
3. Array Color Modulated (*ACM*)
4. Array Size Modulated (*ASM*)

A distinction of the output images of the four modes can be found in this section 4.1.

3.5.2 Enhancement Mode (*facesMode*)

The *facesMode* configuration variable tells the SPV (*see* subsection 3.2.2 for details) which integrated enhancement module to use in order to enhance the output phosphenated images. It also tells the SPV whether to use VJFR or SFR. Thus, the SPV, with the instructions of this configuration variable, can operate in one of the following modes:

1. NOTHING: no enhancements at all are applied to the images.
2. VJFR ROI M: fills the visual field with the Viola Jones face region.
3. SFR ROI M: same as 2, but for the Statistical face region.

4. VJFR ROI C: fills the visual field with the caricatured Viola Jones face region.
5. VJFR ROI HE: fills the visual field with the Viola Jones face region with face-specific histogram equalization (FSHE) applied.
6. SFR ROI HE: same as 5, but for the Statistical face region.
7. VJFR ROI M TD: fills the visual field with the Viola Jones face region of the face with the highest probability of being talking among multiple visible faces according to the Talking Detection module (*see* 3.4.3 for details).
8. SFR ROI M TD: same as 7, but for the Statistical face region.
9. VJFR ROI M ER: fills the visual field with the Viola Jones face region. It utilizes the first three phosphenes from the left in the topmost row to display a binary encoding of the most likely expressed emotion decided by the emotion recognition module (*see* 3.4.2 for details) as shown in Figure 3.16.
10. SFR ROI M ER: same as 9, but for the Statistical face region.

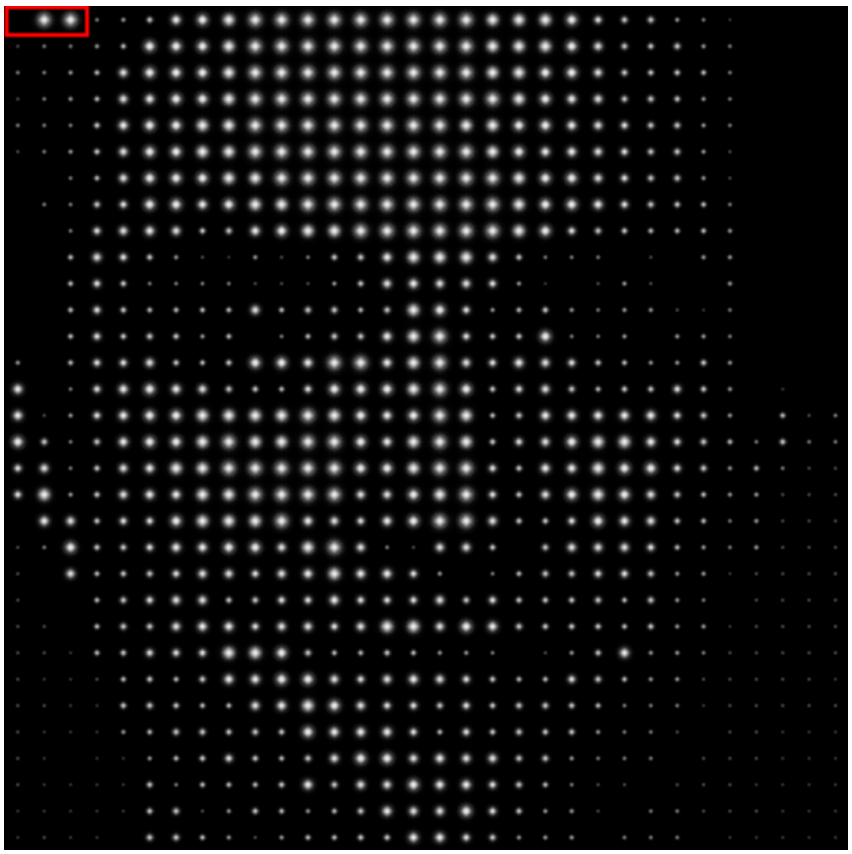


Figure 3.16: The three phosphenes bounded by the drawn rectangle represent a binary encoding of "happy" (011). Each recognizable emotion has a unique encoding.

3.6 Computer Screen Experiment

In this section, the design of the experiment conducted on the computer screen is discussed along with the tests that subjects have undergone. The experiment was conducted on six groups to evaluate the different enhancement modes existing in the SPV. All enhancements have been tested with the Viola Jones face region because even though the Statistical face region shows more promising results in past studies [13] (see subsection 2.6.2 for details), it includes many extra-face information. Extra-face information, like hair, would not be very useful in everyday conversation settings where the facial expressions and facial details are more valuable and should not be compromised for the favor of extra-face information. In addition, some might argue that extra-face details should not be relied on [11] in face recognition (*see* 2.6.3 for more details).

3.6.1 Experiment Groups

There are settings that were common among all groups, and a list of these common settings is presented below:

1. Number of subjects: 5 randomly chosen subjects of both genders at the beginning of their twenties.
2. Experiment device: Lenovo B50-80 with the following specifications:
   - Memory: 11.6 GiB
   - Processor: Intel Core i5-5200U CPU @ 2.20GHz × 4
   - Graphics: AMD Hainan / Mesa Intel HD Graphics 5500 (BDW GT2)
   - OS name: Ubuntu 20.04.4 LTS
   - Screen dimensions:
     - 1366px x 768px
     - 344mm x 193mm
3. SPV configuration (see section 3.1 for more details):
   - *dim*: 32 [32×32 phosphenes]
   - *dimWin*: 640 [$640px \times 640px$ generated phosphenated images]
   - *mLevels*: 16 [modulation levels or colors]
   - *simode*: Array Size Modulated (*ASM*)
4. Field of view: maximum visual angle is 20°. To achieve this, subjects were seated at a distance of $77.4cm$ from the simulation-displaying screen. Check Figure 3.17 for a depiction of experiment settings.

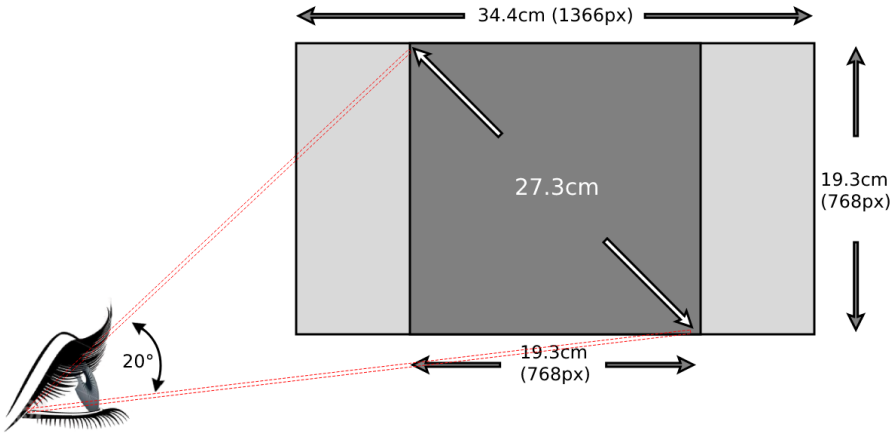


Figure 3.17: Experiment field of view settings. The light gray rectangle is the device screen. The dark grey square on the screen represents the phosphene simulation. Subjects were seated 77.4*cm* away from the screen.

In addition, there are settings that differed among different experiment groups. A list of the groups with their different settings is presented below:

1. Control Group (DLR)
   - *facesMode*: NOTHING
   - The Histogram equalization step in the image simulation preprocessing (see subsection 3.3.1 for details) is skipped
   - Direct lowering of resolution (DLR) is applied
2. Viola Jones face region magnification group (VJm)
   - *facesMode*: VJFR ROI M
   - *ur*: 5
3. Viola Jones FSHE group (VJhe)
   - *facesMode*: VJFR ROI HE
   - *ur*: 5
4. Viola Jones caricaturing group (VJc)
   - *facesMode*: VJFR ROI C
   - *ur*: 1 [faces have to be updated on every frame to apply caricaturing on the proper face bounding boxes]

5. Viola Jones talking detection group (VJtd)
    - *facesMode*: VJFR ROI M TD
    - *ur*: 1 [faces are detected on every frame because the talking detection module needs 25 frames to operate on]
6. Viola Jones emotion recognition group (VJer)
    - *facesMode*: VJFR ROI M ER
    - *ur*: 5 [detected emotion is updated every 5 frames as well]

3.6.2 Experiment Tests

After exposing the subjects to a demo muted phosphenated video, so that they get acclimated to the simulation environment, three tests were conducted. All tests videos were muted because subjects must not be vocally aided. In addition, tests videos were simulated in real-time and played to the subjects. Videos could not be pre-phosphenated (i.e. they had to be simulated in real-time) because subjects must have the liberty to switch between faces if multiple faces are present in the visual field (see section 3.2.2 for more details), and they must have the liberty to skip filling the visual field with a detected face. Each subject sex and age were recorded, and they were randomly assigned to an experiment group. A subject had undergone three tests that are discussed in detail in this section.

1. **Recognition test**
   The goal of this test was to assess the subjects ability in recognizing the identity of a displayed face. The ability of the subjects to see the facial details of the person they see, measured by how accurately would a subject recognize the feeling of the person they see, was assessed as well. The steps of this test are listed below:
    - Subjects were initially shown a long list of 15 Egyptian celebrities. Then, they were asked if they can confidently recognize each face of every listed celebrity. If a subject is not familiar with a celebrity or more in the list, they were familiarized with them by showing the subject their face(s).
    - Each subject then watched a video comprising eight consecutive clips. Each of the 8 clip was showing one of the listed celebrities talking for 14 seconds.
    - For each clip, the subject was asked to:
        (a) Name the talking celebrity.
        (b) Recognize the talking celebrity quickly, for they only had 14 seconds, and their response time was recorded.
        (c) Tell how confident they were of their recognition on a scale of 1 to 5.

(d) Tell whether the talking celebrity is crying. One of the eight clips contained a crying celebrity. The ability of the subjects to capture the crying celebrity was used to assess how clearly were they seeing the facial details of the celebrities. The emotion recognition group (VJer) were taught the binary index of the sad emotion (see subsection 3.5.2 for details), and, therefore, they were expected to have an edge in identifying the crying celebrity. VJer enhancement was not expected to instantaneously help visual prostheses implantees in seeing facial details, but with the help of the neuroplasticity of the brain (see Figure 2.4 for details), the implantees are expected to gradually see facial details more clearly.

2. **Talking test**
   The aim of this test was to assess the subjects ability in identifying the talking person if more than one person were visible in their visual field and only one of them was talking. Each subject watched 4 short video clips, and each video clip showed two people, one on the left and the other on the right, and only one of them was talking. The subjects were asked to identify the talking person position (i.e. left or right), and their answers were recorded. The subjects in groups with enhancements were taught how to switch between faces since only one face would fill the visual field (see subsection 3.2.2 for details). They were also taught how to skip filling the visual field with a face so that they manage to identify the position of the face (i.e. left or right) that they think is talking. The talking detection group (VJtd) subjects were expected to have an edge in this test. A sample screen-shot from the talking test videos is shown in Figure 3.18.

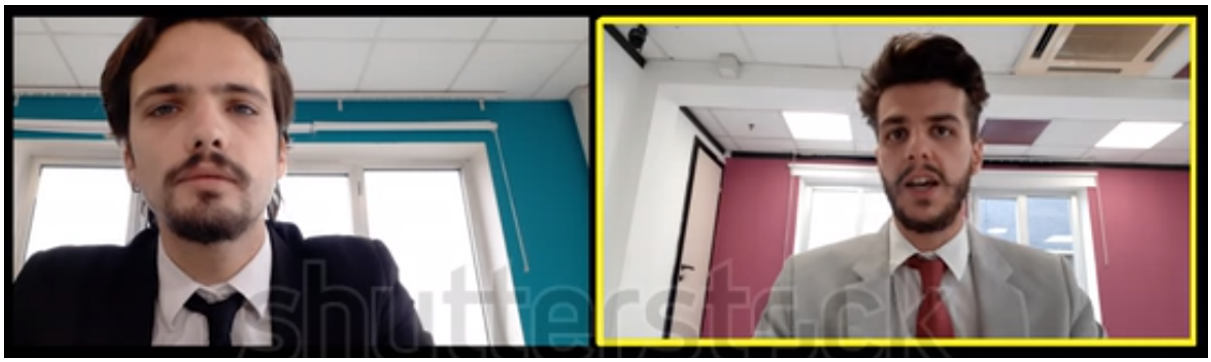


Figure 3.18: A screen-shot from the talking test clips. This clip had the person on the right talking while the one on the left was listening. For a group with ROI magnification applied, the video would typically start with the face of one person filling the visual field. For a subject to identify the talking person position, they would switch between faces multiple times. Once they managed to confidently capture the talking face, they would skip filling the visual field with their face. Skipping the ROI magnification would make a zoom-out effect which was helpful for the subjects in identifying whether that face is located on the left or on the right of the visual field.

3. **Distinction test**
   The goal of this test was to assess the subjects ability in spotting the differences between two face. For this, a video consisting of 7 consecutive clips was played. Each clip was showing a static frame for 10 seconds containing two faces, one on the left and the other on the right. The two faces were either of the same person or of two different people who have common general category (age, sex, or race) and extra-face (hair, or glasses) details. The subjects were then asked if the two displayed images were of the same person or of two different people. Again, the subjects that had enhancements applied were taught how to switch between faces since only one face is allowed to fill the visual field (see subsection 3.2.2 for details). The caricaturing group (VJc) subjects were expected to have an edge in this test. A sample screen-shot from the distinction test videos is shown in Figure 3.19.

Figure 3.19: A screen-shot from the distinction test video clips. This clip has two images of two different people sharing the same general category and extra-face details.

3.7 Virtual Reality Experiment

The computer screen experiment was extended in virtual reality (VR) for a more immersive experience for the subjects. The offered immersiveness was intended to validate the results of the computer screen experiment. This validation was necessary since the computer screen experiment does not isolate the sight of the subjects to make them undergo the tests like actual implantees. In this section, the design of the experiment conducted in VR (with the help of an **Oculus** HMD) is discussed along with the tests that subjects have undergone. The experiment was conducted on three groups to validate the ROI magnification and caricaturing impacts on the subjects performance achieved in the computer screen experiment (see section 4.2 for details).

3.7.1 Experiment Groups

There are settings that were common among all VR groups:

1. Number of subjects: 5 randomly chosen subjects of both genders at the beginning of their twenties.
2. Experiment device: MSI GF63 with the following specifications:
   - Memory: 16 GiB
   - Processor: Intel Core i7-11800H 8 Core 2.4-4.6 GHz
   - Graphics: NVIDIA GeForce RTX3050 Laptop GPU — 4G GDDR6
   - OS name: Windows 10
   - Screen dimensions:
     - 1920px x 1080px or 349mm x 196mm
3. HMD: **Oculus Quest 2**
4. SPV configuration (see section 3.1 for more details):
   - *dim*: 32 [32×32 phosphenes]
   - *dimWin*: 640 [$640px \times 640px$ generated phosphenated images]
   - *mLevels*: 16 [modulation levels or colors]
   - *simode*: Array Color Modulated (*ACM*) [the *simode* was initially set to ASM, but the subjects were unable to see any meaningful details even with enhancements applied, so it was changed to ACM to have varying meaningful results]

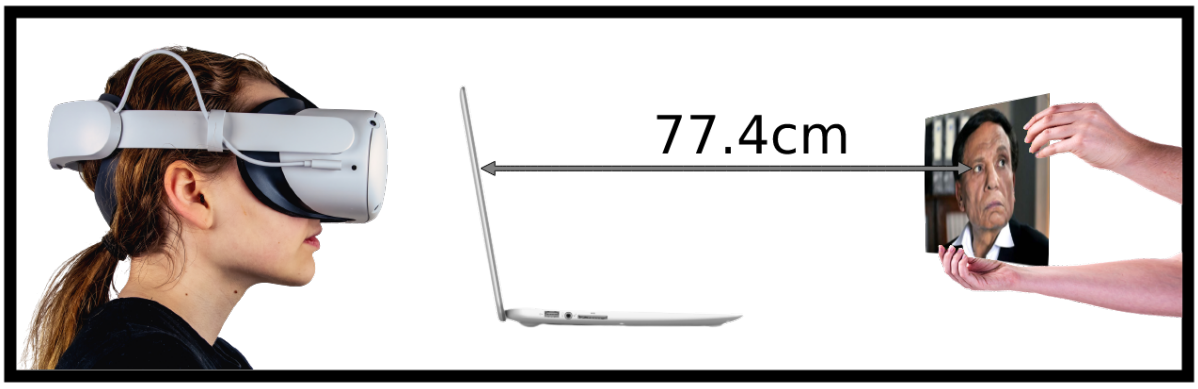


Figure 3.20: In the VR experiment, subjects were watching the experiment device desktop virtually inside the VR. The desktop of the experiment device was displaying the scene captured by its web-cam after being phosphenated. Test images were held still at a distance 77.4cm from the experiment device web-cam.

5. Experiment setup: Subjects were asked to wear the **Oculus** HMD. In the VR, a virtual desktop (as shown in Figure 3.21) was displayed at a distance of 2.7m. The desktop was that of the device used to run the SPV. The web-cam of the device was used to capture the surrounding scene, which was then phosphenated and displayed on the device desktop; the desktop that was displayed virtually inside the **Oculus** VR. The test images were held still for the subjects at a distance of 77.4cm from the device web-cam (as shown in Figure 3.20)

Figure 3.21: Oculus Virtual Desktop.

In addition, there are settings that differed among different experiment groups:

1. Control Group (DLR)
    - *facesMode*: NOTHING
    - The Histogram equalization step in the image simulation preprocessing (see subsection 3.3.1 for details) is skipped
    - Direct lowering of resolution (DLR) is applied
2. Viola Jones face region magnification group (VJm)
    - *facesMode*: VJFR ROI M
    - *ur*: 5
3. Viola Jones caricaturing group (VJc)
    - *facesMode*: VJFR ROI C
    - *ur*: 1 [faces have to be updated on every frame to apply caricaturing on the proper face bounding boxes]

3.7.2 Experiment Tests

To make the subjects familiar with the experiment environment, they were first left to explore the displayed scene for a short period of time. The experiment device web-cam would often capture the experiment conductor face, and the subjects were asked whether they can see the face clearly or not. After that, three tests were conducted. Conducted tests relied on capturing the surrounding scene and phosphenating it in real-time for the VR subjects. Subjects also had the liberty to switch between faces if multiple faces are present in the visual field (see section 3.2.2 for more details), and they had the liberty to skip filling the visual field with a detected face. Each subject sex and age were recorded, and they were randomly assigned to an experiment group. A subject had undergone three tests that are discussed in detail in this section.

1. **Recognition test**
   The goal of this test was to assess the subjects ability in recognizing the identity of a displayed face. The steps of this test are listed below:

   - Subjects were initially shown a long list of 15 Egyptian celebrities. Then, they were asked if they can confidently recognize each face of every listed celebrity. If a subject is not familiar with a celebrity or more in the list, they were familiarized with them by showing the subject their face(s).
   - A printed image of one of the listed celebrities or of an unknown person (to which the subjects were expected to say, "I do not know," and it would have been counted as a correct response) would then be held in front of the experiment device (as shown in Figure 3.20) web-cam for 10 seconds. This was repeated for 6 different images.
   - For each image, the subject was asked to
     (a) Name the shown celebrity, or say I do not know.
     (b) Respond quickly to question (a), for they only had 10 seconds, and their response time was recorded.
     (c) Tell how confident they were of their recognition on a scale of 1 to 5.

2. **Expressions test**
   The aim of this test was to assess the subjects ability in seeing the facial details of a face existing in their visual field. Each subject was first familiarized with 4 different expressions on the face of the experiment conductor. The expressions were happy, sad, frowning, and surprised (as shown in Figure 3.22). The experiment conductor would then mimic 4 of these expressions (P.S. an expression could be repeated) using his face in front of the experiment device web-cam at a distance of 77.4cm (see Figure 3.20 for a depiction of the experiment setup), and the subjects would then be asked to identify the expression expressed. The experiment conductor would express an expression for 10 seconds, and the subjects were expected to answer within this time frame.

Figure 3.22: Different expressions that were shown to the subjects during the expressions test on the face of the experiment conductor.

3. **Distinction test**
   The goal of this test was to assess the subjects ability in spotting the differences between two face. For this, printed images containing two faces, one on the left and the other on the right, were shown to the subjects through the SPV device web-cam (see Figure 3.20 for details). The two faces were either of the same person or of two different people who have common general category (age, sex, or race) and extra-face (hair, or glasses) details. The subjects were then asked if the two displayed images were of the same person or of two different people. In this test, 4 different images were displayed to the subjects. The subjects that had enhancements applied were taught how to switch between faces since only one face is allowed to fill the visual field (see subsection 3.2.2 for details).

Chapter 4

Results

After the SPV was implemented, it was published on a public **GitHub** repository [16]. The repository has a **read-me** file that describes in detail how to clone and run the SPV. The **read-me** file also details the repository content: the repository contains the test videos that were used in the computer screen experiment tests along with the SPV itself.

This chapter discusses the results of the both the phosphenes simulation (i.e. the generated phosphenated images) and the conducted experiments. These results could be validated, or ideally replicated, with the help of the SPV **GitHub** repository found in this url: https://github.com/HeshamMoneer/Phosphenes-Simulation.

4.1 Simulation output images

As mentioned in subsection 3.5.1, there are four possible ways for the SPV to draw a phosphene. A distinction between the output phosphenated images of the four ways is shown in Figure 4.2. According to the literature [5], Gaussian size modulated phosphenes are closer to the actual shape of physiological phosphenes. Accordingly, it was intended to use size modulated phosphene in both computer screen and virtual reality experiments. The size modulated phosphenes in the virtual reality experiment, however, made the subjects unable to see any details even with enhancements applied. Therefore, the size modulated phosphenes were only used in the computer screen experiment, and color modulated phosphenes were used in the virtual reality experiment.

The preferred method to draw phosphenes (see subsection 3.5.1 for details) was the array of circles method. The phosphenes drawn with this method were more like Gaussian profiles than phosphenes drawn with the Blur method. The Figure 4.1 shows what it would look like if the same pixel was phosphenated using the four different phosphenes drawing methods.

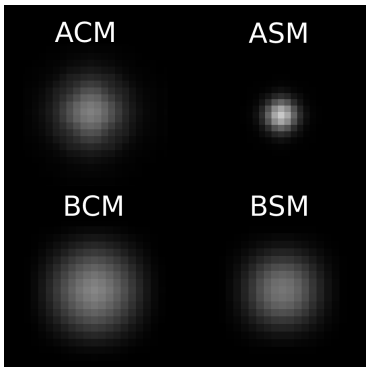


Figure 4.1: This figure shows the four different possible ways the implemented SPV can phosphenate a pixel. These four phosphenes are of the same input pixel but phosphenated using the four different methods.

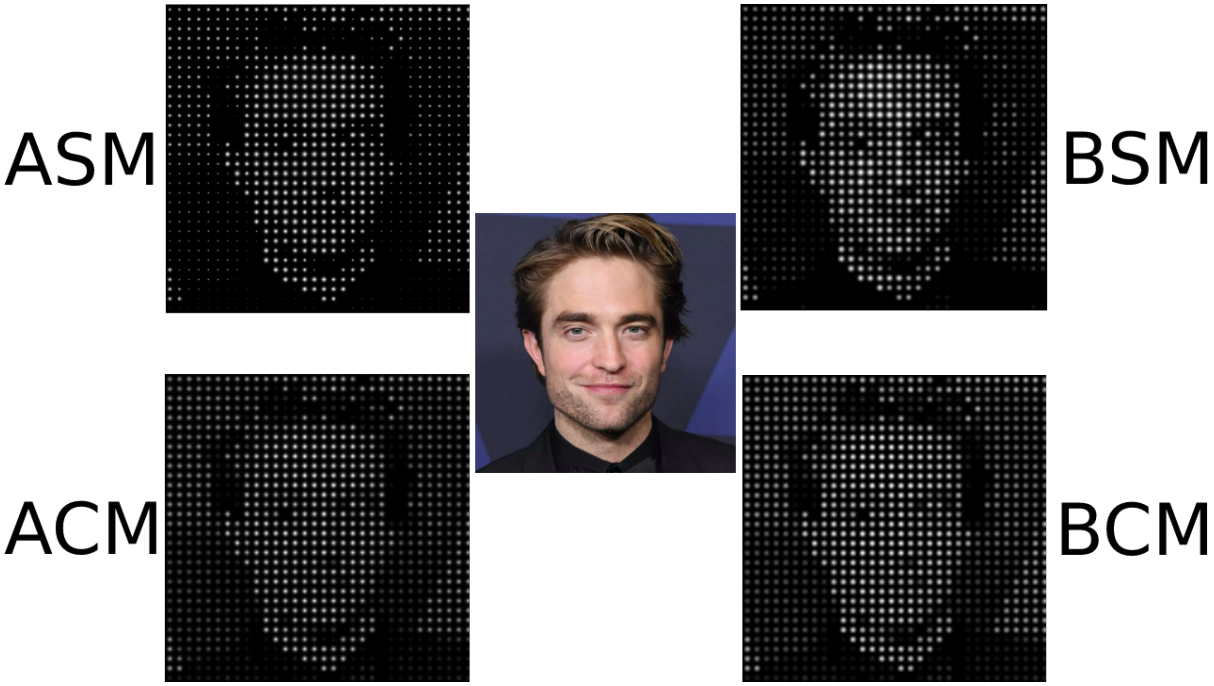


Figure 4.2: This figure shows the four different possible ways the implemented SPV can phosphenate a source image. The input source image is the one in the middle of the figure, and the other four peripheral images are output phosphenated images labeled by which method their phosphenes were drawn.

4.2 Computer Screen Experiment

In this section, the computer screen experiment results are shown and discussed. The experiment results (i.e. subjects demographics and performance in various conducted tests) are shown in table 4.1, and they are also summarized as bar charts with error bars (see Figures 4.3, 4.4, 4.5, 4.6, 4.7).

In table 4.1, the No. subjects who detected the crying person row is the number of the subjects who managed to detect the crying person as the only person crying (i.e. the responses of the subjects who detected more than a crying person, no crying people, or a wrong crying person were counted as wrong responses). The majority of the subjects were Computer Engineering students in their fourth year, which explains the dominance of males and the 21 year olds over the experiments participants.

	DLR	VJm	VJc	VJtd	VJer	VJhe
Genders	5m + 0f	5m + 0f	3m + 2f	4m + 1f	3m + 2f	4m + 1f
Average age	21.6 y/o	21.4 y/o	22 y/o	20.6 y/o	21.2 y/o	21 y/o
Recognition Test Results						
Accuracy percentages	62.5% 25% 50% 0% 0%	75% 50% 75% 75% 62.5%	75% 62.5% 100% 87.5% 100%	87.5% 75% 62.5% 100% 75%	100% 75% 37.5% 75% 100%	87.5% 62.5% 50% 75% 75%
Average response times	3.5sec 3.4sec 4.1sec 7sec 3.5sec	3.1sec 5.1sec 7sec 4.1sec 4.1sec	5.5sec 5.8sec 2.87sec 4.1sec 5.1sec	3.1sec 3.3sec 3.3sec 1.8sec 3.1sec	2.5sec 7.4sec 9sec 3.6sec 3.6sec	1.1sec 3.4sec 4.3sec 3.1sec 2.6sec
Average confidence levels	4 2.25 2.63 0.6 0.75	4 2.3 3.25 3.25 3.25	4 3.87 4.87 4 3.8	4.5 4 2.6 5 3	4.63 3.8 2.6 4.3 5	4.4 3 3.25 4 4
No. subjects who detected the crying person	0	0	1	2	2	0
Talking Test Results						
Accuracy percentages	50% 100% 0% 75% 75%	100% 50% 25% 100% 100%	0% 50% 75% 75% 100%	25% 50% 50% 50% 25%	100% 75% 100% 100% 100%	75% 75% 50% 75% 25%
Distinction Test Results						
Accuracy percentages	42.8% 28.7% 14.3% 57.1% 0%	42.8% 42.8% 28.5% 85.7% 57.1%	71.4% 57.1% 85.7% 42.8% 42.8%	57.1% 71.4% 14.3% 71.4% 71.4%	42.8% 71.4% 57.1% 42.8% 57.1%	71.4% 42.8% 28.6% 28.6% 28.6%

Table 4.1: Computer screen experiment results.

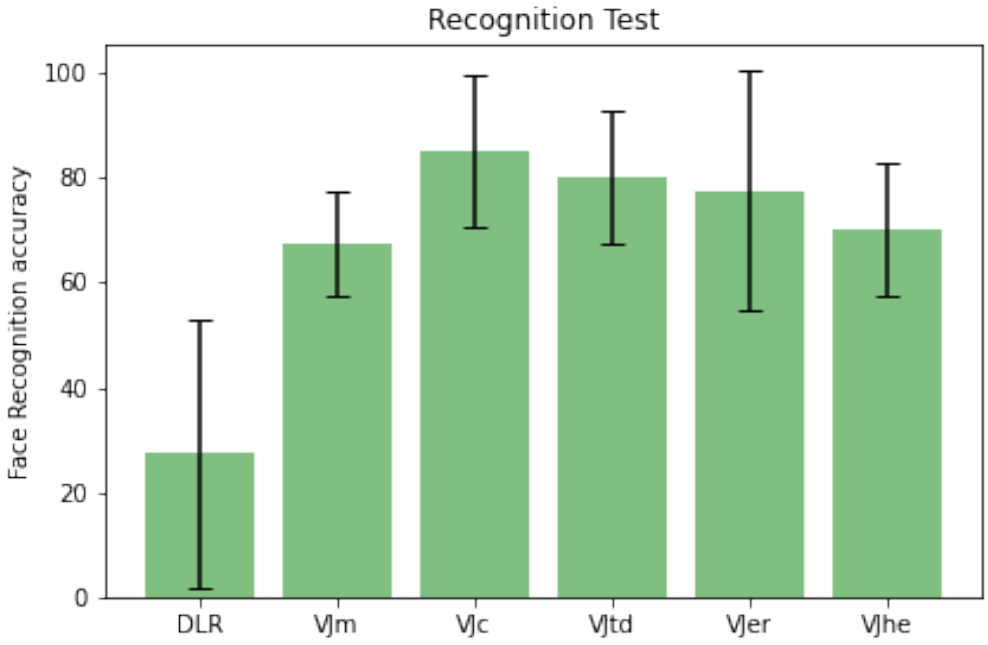


Figure 4.3: A comparison between the accuracies achieved by subjects in different computer screen experiment groups in recognizing faces in recognition test.

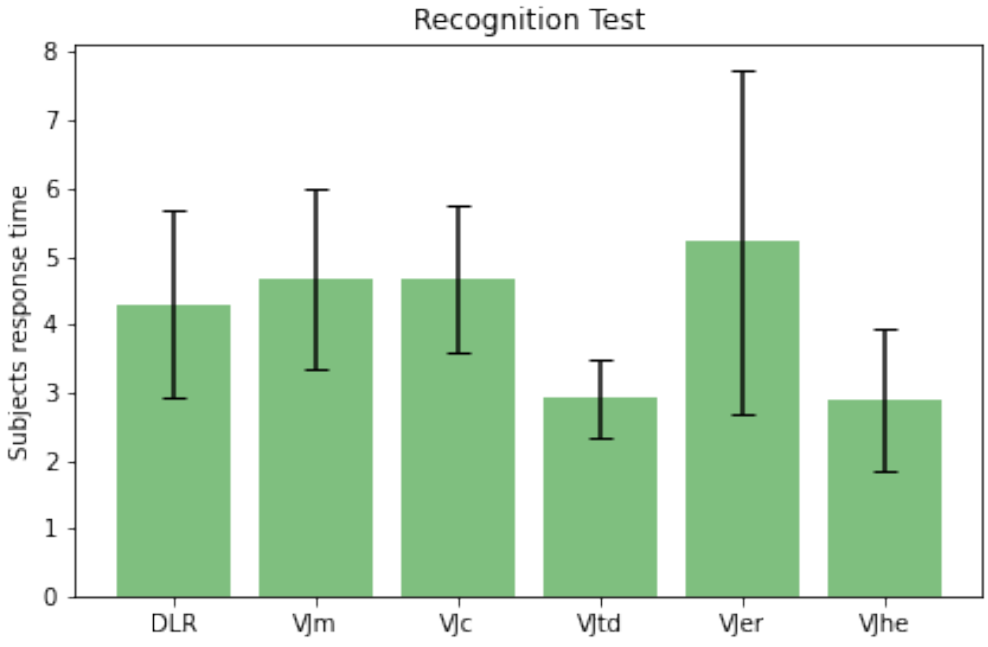


Figure 4.4: A comparison between the response times taken by subjects in different computer screen experiment groups to recognize faces in recognition test.

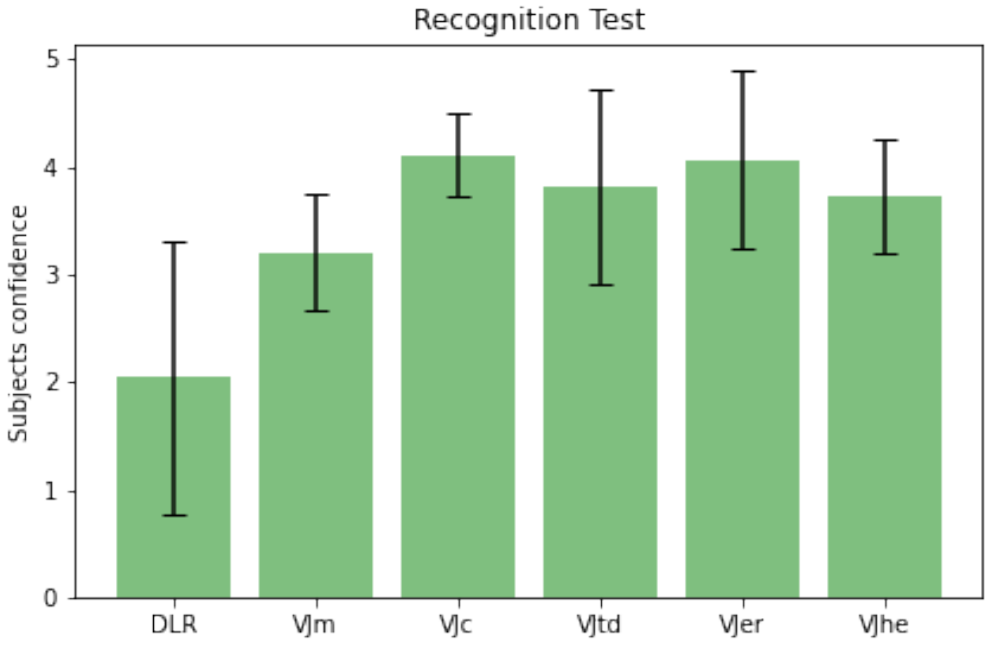


Figure 4.5: A comparison between the confidence levels reported by subjects in different computer screen experiment groups while recognizing faces in recognition test.

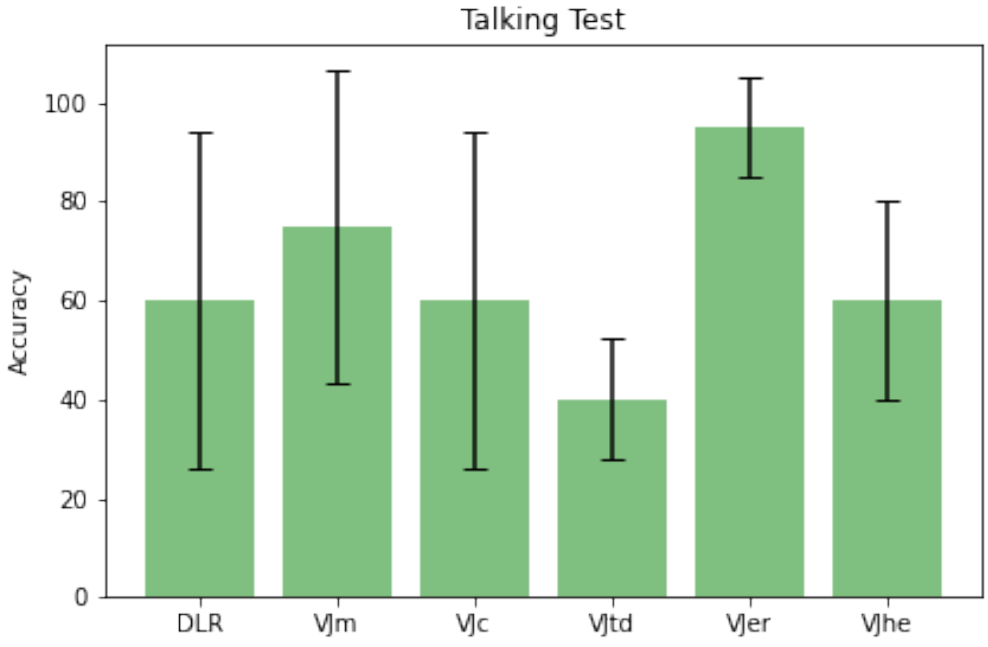


Figure 4.6: A comparison between the accuracies achieved by subjects in different computer screen experiment groups in detecting the talking person position in talking test.

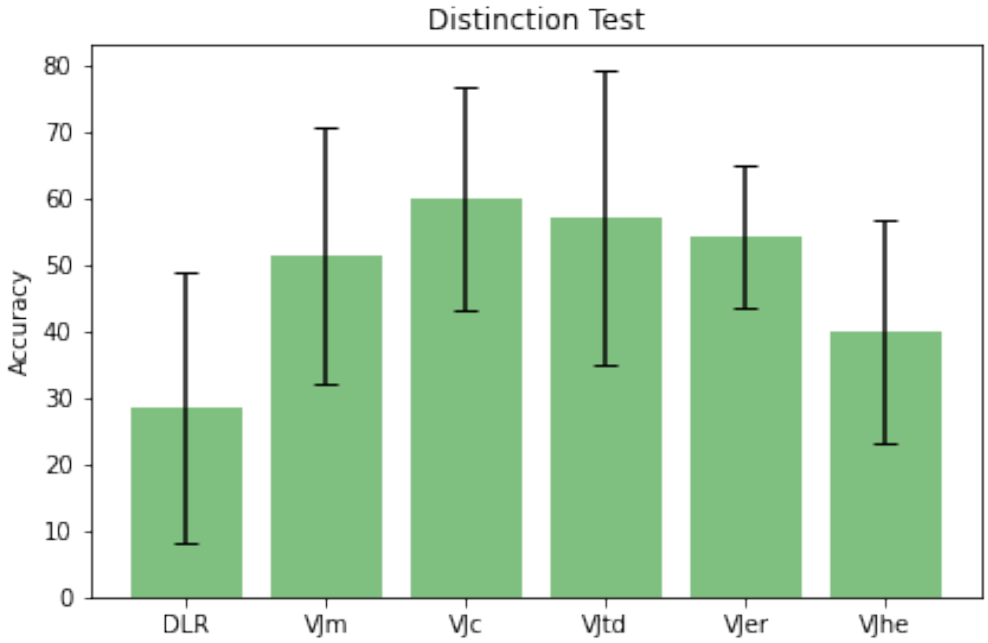


Figure 4.7: A comparison between the accuracies achieved by subjects in different computer screen experiment groups in distinguishing faces in distinction test.

The bar charts drawn in Figures 4.3 to 4.7 represent visualizations of the data listed in the table 4.1, which are the results of the computer screen experiment. The bar of a test group is a visualization of their mean value, and the error bars are visualizations of each group standard deviation. The difference between two groups means is statistically significant if their error bars do not overlap.

The results visualization suggest that Viola Jones ROI magnification applied in groups other than the control group made a significant difference in increasing the subjects face recognition ability (as shown in Figure 4.3). ROI magnification, however, did not significantly affect the subjects response time (as shown in Figure 4.4) or confidence (as shown in Figure 4.5). The tested enhancement modules in this thesis did not show significant results in achieving their expected goals as discussed below:

- Caricaturing module:
  The caricaturing group was expected to show significantly higher accuracy in distinction test. The results visualized in Figure 4.7 suggest that only insignificant higher accuracy was achieved by the caricaturing group when compared to the other groups.

- Talking detection module:
  The talking detection group was expected to show significantly higher accuracy in talking test. The results visualized in Figure 4.6 suggest that the module was actually hindering the subjects from detecting the talking person correctly. This mostly

happened because the module recurrent neural network (RNN) relied heavily on the detected face landmarks, which relied heavily on the Viola Jones Haar cascade. This led to an over-accumulation of error, which made the module decision inaccurate. In addition, the magnified ROI would switch every time a face with higher probability of being talking is detected depriving the subjects from having absolute control on which face to magnify.

- Emotion recognition module:
  The emotion recognition module was expected to increase the number of the subjects that are able to detect the crying person. The collected data in table 4.1 suggest that it did not significantly increase the number of subjects able to detect the crying person. This mostly happened because the sbujects had to focus on so many details (i.e. the facial details of the person they see, and the emotion binary index). Also, an emotion binary index that would update once every 5 frames (see subsection 3.6.1 for details) was definitely distracting the subjects.

- Face-Specific Histogram Equalization module:
  The FSHE group was expected to show a significantly higher face recognition accuracy in the recognition test. The results visualized in Figure 4.3 suggest that the FSHE did not achieve the desired higher face recognition accuracy.

The caricaturing module has shown only insignificant difference in improving the subjects performance in the distinction test. This might have happened due to the small number of subjects who had undergone the tests within each group and the limited number of questions they were asked.

The virtual reality extension of the experiment was, therefore, conducted on only three groups, namely DLR, VJm, and VJc. The purpose of this was to validate the ROI magnification significantly higher subjects face recognition accuracy achieved, and to further experiment the effect of face caricaturing on increasing the subjects distinguishability of faces and ability to see facial details.

4.3 Virtual Reality Experiment

In this section, the virtual reality experiment results are shown and discussed. The experiment results (i.e. subjects demographics and performance in various conducted tests) are shown in table 4.2, and they are also summarized as bar charts with error bars in Figures 4.8, 4.9, 4.10, 4.11, and 4.12.

The results confirm that ROI magnification actually significantly improves subjects accuracy in recognizing faces (see Figure 4.8). The results also show that ROI magnification significantly improves subjects response time and confidence in recognition test (see Figures 4.9 and 4.10), subjects accuracy in capturing facial expressions (see Figure 4.11), and subjects accuracy in distinguishing faces (see Figure 4.12).

In the distinction test, the caricaturing group was expected to show higher performance when compared to the normal ROI magnification group. This might have not been achieved (see Figure 4.12) due to a lot of reasons including the low quality caricaturing algorithm implemented, the small number of experiment participants and the limited number of pictures they were shown, or the unfamiliarity of the subjects with the VR environment.

The majority of the subjects were Computer Engineering students in their fourth year, which explains the dominance of males and the 21 year olds over the experiments participants.

	DLR	VJm	VJc
Genders	4m + 1f	4m + 1f	3m + 2f
Average age	21.6 y/o	21.6 y/o	21.4 y/o
Recognition Test Results			
Accuracy percentages	33.3% 33.3% 0% 0% 0%	66.6% 33.3% 100% 83.3% 83.3%	83.3% 83.3% 100% 100% 83.3%
Average response times	4sec 7.3sec 10sec 8.8sec 10sec	4sec 8.3sec 2.3sec 3sec 2.8sec	4.1sec 4.3sec 2.6sec 3.2sec 4.6sec
Average confidence levels	2.6 1.8 0 0.3 0	4.3 3.3 4.8 4.2 4.8	4 4.1 4 4 3.5
Expressions Test Results			
Accuracy percentages	0% 0% 0% 0% 0%	100% 100% 50% 100% 100%	25% 100% 100% 50% 100%
Distinction Test Results			
Accuracy percentages	50% 50% 0% 25% 0%	100% 75% 100% 25% 75%	50% 25% 50% 75% 50%

Table 4.2: Virtual Reality experiment results.

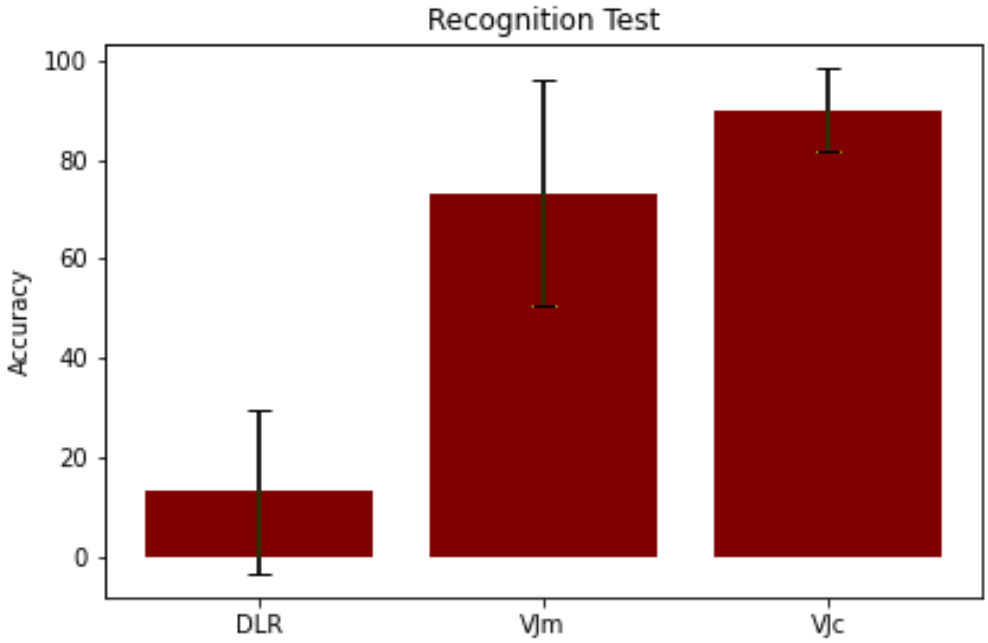


Figure 4.8: A comparison between the accuracies achieved by subjects in different virtual reality experiment groups in recognizing faces in recognition test.

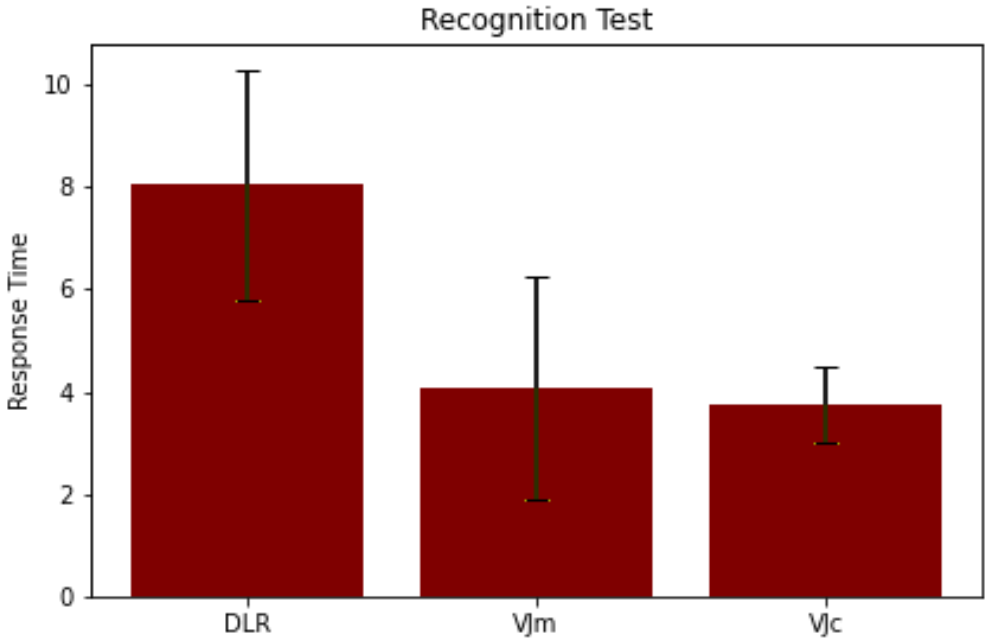


Figure 4.9: A comparison between the response times taken by subjects in different virtual reality experiment groups to recognize faces in recognition test.

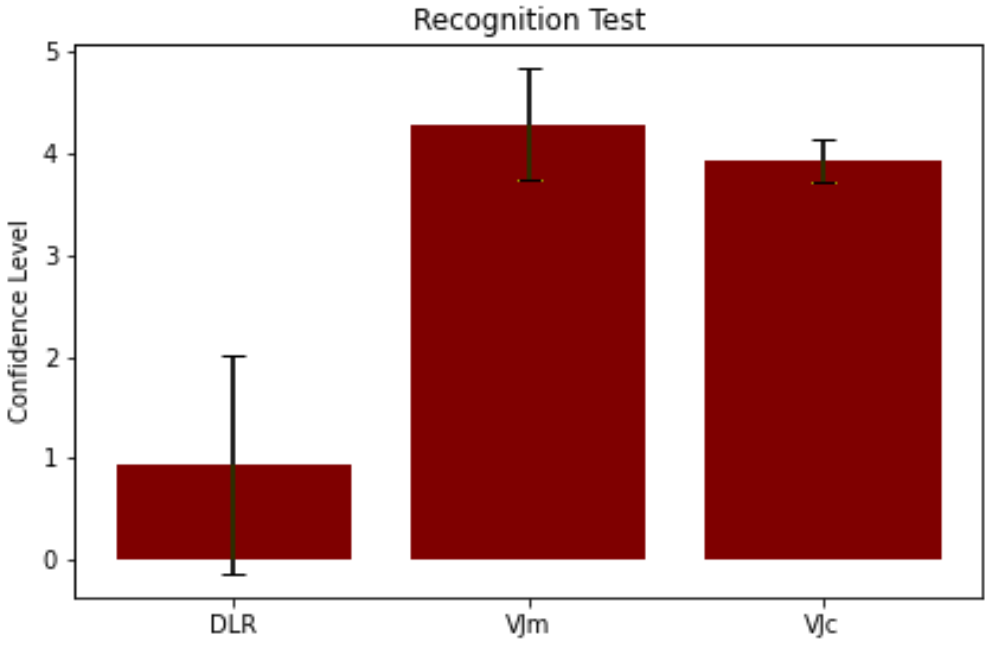


Figure 4.10: A comparison between the confidence levels reported by subjects in different virtual reality experiment groups while recognizing faces in recognition test.

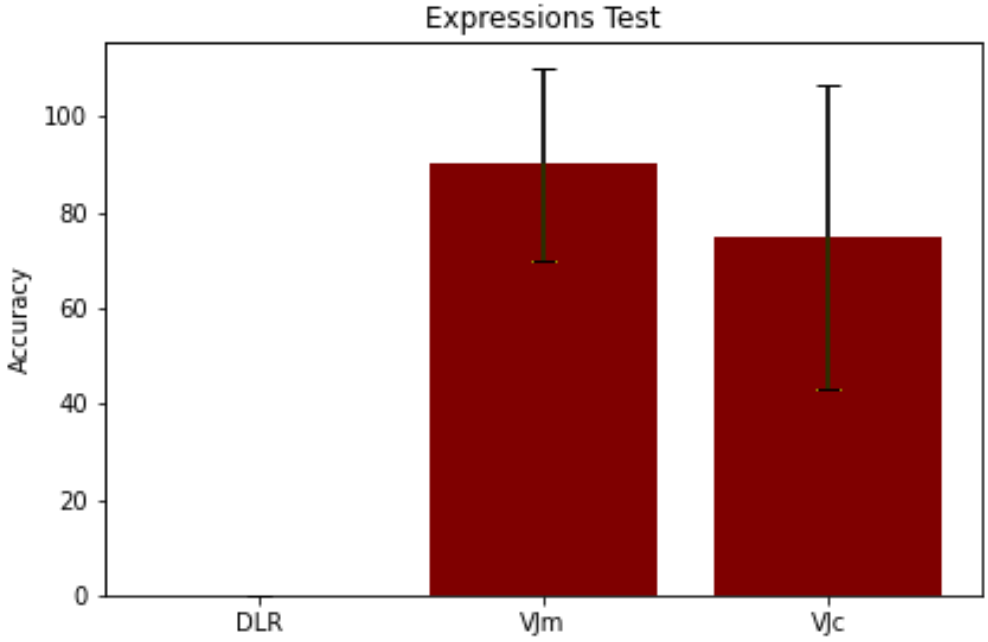


Figure 4.11: A comparison between the accuracies achieved by subjects in different virtual reality experiment groups in detecting the experiment conductor expressions in expressions test.

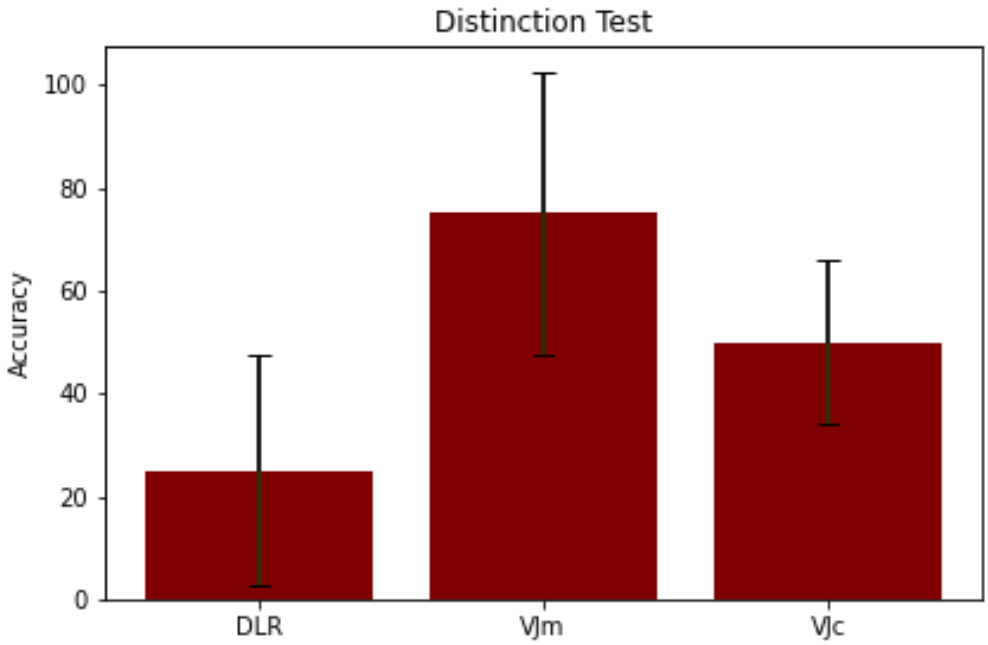


Figure 4.12: A comparison between the accuracies achieved by subjects in different virtual reality experiment groups in distinguishing faces in distinction test.

4.4 Applicability in real-time

In order to test whether the algorithms implemented in different enhancement modules are applicable in real-time, two measures were taken. First, the phosphenated scene was checked for any lag every time the SPV had to run. No lag was ever observed by the experiments conductor or the subjects undergoing the experiments in neither the computer screen experiment nor the virtual reality experiment. The second measure was to calculate how long did it take the SPV to play a video in phosphenes. That runtime was then compared with the actual duration of that video, and the difference never exceeded a fraction of a second. The second measure was taken only in the computer screen experiment. It was not possible to take the second measure in the virtual reality experiment since the surrounding scene was captured, and the tests runtime would have had no reference to be compared with.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, multiple options (options that are applications of image processing, computer vision, and machine learning techniques) were explored to enhance visual prostheses implantees ability to recognize faces and see facial details. These options are discussed in the subsequent paragraphs.

An attempted approach was to fill the visual field with a detected face: this way, the limited number of phosphenes in the visual field would then convey only useful information of facial details.

Another approach was caricature a detected face: caricaturing would exaggerate facial features which would in turn help the implantees to see these features more clearly.

The third approach was to highlight the emotion expressed by a detected face: highlighting the emotion expressed by a detected face would help the implantees in guessing the expression of the face they see. With suitable learning and rehabilitation, implantees would use the highlighted emotions to see the facial details more clearly.

The fourth approach was to detect the talking person: the implantees would have the opportunity to focus on the face of the person talking, which would help them in a group conversation setting.

The last explored approach was face-specific histogram equalization (FSHE): Histogram equalization (improved contrast) applied to the face region would help overcome any unclarities in the facial details caused by poor lighting conditions. This would help the implantees in seeing facial details more clearly.

To compare the impacts of the proposed options on the visual prostheses implantees performance, a Simulation of Prosthetic Vision (SPV) was implemented and was experimented on normally sighted subjects in two experiments. The first experiment was conducted on a computer screen and the second experiment was conducted in virtual

reality (VR) using an **Oculus** head-mounted display (HMD). The results of the conducted experiments suggest that filling the visual field with a detected face (Viola Jones region-of-interest (ROI) magnification) has a significant impact on the subjects ability to recognize faces and see facial details.

Although Caricaturing a magnified face did not significantly improve subjects performance, when compared to mere magnification applied, magnification with caricaturing is concluded to be the best option for enhancing faces because exaggerated facial features were more evident for the subjects.

5.2 Limitations and Future Work

Many limitations were faced in this thesis. These limitations include the following:

1. The number of experiments participants and the number of test images and videos they were shown were limited due to the limited time allowed for the experiments.

2. The experiments were conducted on normally sighted subjects using a simulation because actual implantees were not accessible.

3. The caricaturing algorithm was of relatively low quality. The quality of the module was lowered on purpose in order to make the module run in real-time settings.

4. Facial detection and magnification applied on consecutive video frames, on either a prerecorded video or a web-cam captured real-time video, resulted in a very unstable visual field for the subjects even with a more idle detection rate of faces (see subsection 3.2.1 for details). For example, a nodding or a shaking face would cause the visual field to update so frequently causing the field to be unstable, which was an inconvenient experience for the subjects. Another example is the following: Assume a visual field that contains a face that is not clearly visible. If that face is to be captured for three consecutive frames, and the Viola Jones Haar Cascade Classifier manages to detect it in the first and the last frame but not in the middle, this would result in great instability of the visual field.

5. Highlighting the face emotion in the visual field of the subjects has proven to be distracting for the subjects.

6. Detecting the talking person deprived the subjects from having utter control on which face to magnify. This also caused great instability to the visual field because if more than one face are present in the visual field, the talking detection module would repeatedly swap among them based on the decision it makes every time; a decision that suffers from accumulated inaccuracy due to the module recurrent neural network (RNN) great dependence on the detected face landmarks.

Due to these limitations, the work that has been described in this thesis could be extended in so many ways.

First, the experiments could be conducted on a larger number of subjects with more test images and videos included. In addition, experiments results would be of more value if they were tested on actual implantees.

Second, the impact of caricaturing a magnified face could be compared to mere magnification in a separate experiment. Moreover, a more sophisticated caricaturing algorithm could be tested, while preserving the real-time applicability of the algorithm. One possible way to do so is to make use of GANs like CariGAN [15].

At last, the used modules of face-specific histogram equalization (FSHE), talking detection (TD), and emotion recognition (ER) could be improved (e.g. the detected emotions could be highlighted in a non-distracting way, or the talking person could be detected without taking away the subjects control over which face to magnify) and tested in separate experiments because their impacts were not sufficiently explored in this thesis.

Appendix

Appendix A

Lists

RP	retinitis pigmentosa
AMD	age-related macular degeneration
ML	machine learning
CV	computer vision
DLR	directly lower resolution
VJm	Viola Jones face magnification
VJc	Viola Jones face caricaturing
VJtd	Viola Jones talking detection
VJer	Viola Jones emotion recognition
VJhe	Viola Jones face-specific histogram equalization
LGN	lateral geniculate nucleus of the thalamus
SPV	Simulation of Prosthetic Vision
PMI	phosphene modulation index
HMD	head-mounted display
ROI	region-of-interest
DLR	directly lower resolution
VJFR	Viola Jones face region
SFR	Statistical face region
MFR	Matting face region
NPD	nose position detection
API	application programming interface
BCM	blur color modulated
BSM	blur size modulated

ACM	array color modulated
ASM	array size modulated
HE	histogram equalization
FSHE	face-specific histogram equalization
ER	emotion recognition
TD	talking detection
VR	virtual reality
RNN	recurrent neural network
CNN	convolutional neural network
GAN	generative adversarial network

List of Figures

2.1 Phosphenes shape 3
2.2 The Visual Pathway 4
2.3 Main approaches of visual prosthesis. 5
2.4 Neuroplasticity 5
2.5 Visual Prosthesis Design Thrusts 6
2.6 Simulations of Prosthetic Vision 7
2.7 Gaussian Distribution shape 8
2.8 Visual Angle 9
2.9 Visual field lattice shape 10
2.10 Distinguishing light sources 10
2.11 Image pyramid 12
2.12 Integral Image 12
2.13 Rectangular features examples 12
2.14 Haar cascade 13
2.15 Viola Jones algorithm output sample 13
2.16 Face Landmark Detection 14
2.17 Theory of caricaturing and face-space 15
2.18 Face Morphing 15
2.19 Histogram Equalization 16
2.20 ROIs magnification in SPVs 18

3.1 Simulation flowchart 20
3.2 Haar cascade sample ouput 22
3.3 Detecting multiple faces in the visual field 23

3.4 Methods of squaring an image 24
3.5 Drawing Phosphenes 26
3.6 Phosphene Cache 26
3.7 Dlib Face Landmarks 27
3.8 Face-specific histogram equalization 28
3.9 Emotion Recognition 29
3.10 Talking Detection 30
3.11 Delauny Triangulation 31
3.12 Straying face landmarks 32
3.13 Caricaturing module working steps 33
3.14 Gaussian Convolution Kernel 34
3.15 Phosphene Array drawing method 35
3.16 Emotion recognition module applied on SPV 36
3.17 Computer Screen Experiment setup 38
3.18 Talking test 40
3.19 Distinction test 41
3.20 VR Experiment setup 42
3.21 Oculus Virtual Desktop 43
3.22 VR Expressions Test 45

4.1 Different SPV Phosphenes 47
4.2 SPV Output sample 47
4.3 Computer Screen Recognition Test Accuracies. 49
4.4 Computer Screen Recognition Test Response Times. 49
4.5 Computer Screen Recognition Test Confidence Levels. 50
4.6 Computer Screen Talking Test Accuracies. 50
4.7 Computer Screen Distinction Test Accuracies. 51
4.8 Virtual Reality Recognition Test Accuracies. 54
4.9 Virtual Reality Recognition Test Response Times. 54
4.10 Virtual Reality Recognition Test Confidence Levels. 55
4.11 Virtual Reality Expressions Test Accuracies. 55
4.12 Virtual Reality Distinction Test Accuracies. 56

Bibliography

[1] 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.

[2] E. Fernandez, "Development of visual neuroprostheses: trends and challenges," *Bioelectronic medicine*, vol. 4, no. 1, pp. 1–8, 2018.

[3] I. Bókkon, "Phosphene phenomenon: a new concept," *BioSystems*, vol. 92, no. 2, pp. 168–174, 2008.

[4] E. Messina and J. Evans, "Standards for visual acuity," *National Institute for Standards and Technology*, 2006.

[5] 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.

[6] X. Zhang, *Gaussian Distribution*, pp. 425–428. Boston, MA: Springer US, 2010. https://doi.org/10.1007/978-0-387-30164-8_323.

[7] J. Swearer, *Visual Angle*, pp. 2626–2627. New York, NY: Springer New York, 2011. https://doi.org/10.1007/978-0-387-79948-3_1411.

[8] E. Ackerman, L. B. Ellis, and L. E. Williams, *Biophysical science*, pp. 31–33. Prentice-Hall, 1979.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, Ieee, 2001.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[11] 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.

[12] 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.

[13] 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.

[14] S. S. Agaian, B. Silver, and K. A. Panetta, "Transform coefficient histogram-based image enhancement algorithms using contrast entropy," *IEEE transactions on image processing*, vol. 16, no. 3, pp. 741–758, 2007.

[15] K. Cao, J. Liao, and L. Yuan, "Carigans: Unpaired photo-to-caricature translation," *arXiv preprint arXiv:1811.00222*, 2018.

[16] H. Moneer, "Enhanced Simulation of Prosthetic Vision." https://github.com/HeshamMoneer/Phosphenes-Simulation.

[17] G. Leach, "Improving worst-case optimal delaunay triangulation algorithms," in *4th Canadian Conference on Computational Geometry*, vol. 2, p. 15, Citeseer, 1992.

[18] E. Davies, "Machine vision: Theory, algorithms and practicalities," pp. 42–44, 1990.