

RADBOUD UNIVERSITY NIJMEGEN



FACULTY OF SOCIAL SCIENCES

---

# Phosphene Vision and AI: Representing the World

---

THESIS BSc ARTIFICIAL INTELLIGENCE

*Author:*  
Codruta LUGOJ

*Supervisor:*  
Dr. Yağmur GÜÇLÜTÜRK

July 2019

## **Abstract**

Prosthetic vision relies on the observation that stimulating the visual cortex generates small visual percepts called phosphenes. Due to biological limitations, current cortical implants have low resolution. Researchers started investigating more complex image processing techniques in an attempt to circumvent these limitations and improve the quality of the phosphene image. Semantic segmentation has been proposed as a potential solution to this issue. However, these complex algorithms are computationally demanding. We investigated whether there are any advantages in using complex image processing techniques such as semantic segmentation over simpler methods, here represented by edge detection. We compared the two methods through a visual search task in simulated phosphene vision on healthy subjects, measuring reaction times as well as subjective opinions. Our results show that participants performed better under the edge detection condition. The current state of the semantic segmentation model used in this work needs to be improved before it can yield better results than edge detection.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Methods</b>	<b>5</b>
2.1	Semantic segmentation model . . . . .	5
2.1.1	Base network . . . . .	5
2.1.2	Indoor data and augmentation . . . . .	6
2.1.3	Finetuning . . . . .	6
2.2	Experiment . . . . .	7
2.2.1	Participants . . . . .	7
2.2.2	Setup and task . . . . .	7
2.2.3	Stimuli . . . . .	8
<b>3</b>	<b>Results</b>	<b>11</b>
<b>4</b>	<b>Discussion</b>	<b>14</b>
4.1	Limitations . . . . .	15
<b>5</b>	<b>Future work</b>	<b>16</b>
<b>6</b>	<b>Conclusion</b>	<b>16</b>
<b>7</b>	<b>Appendix</b>	<b>19</b>

# 1 Introduction

Visual prostheses have been proposed as candidates for restoring limited levels of vision in visually impaired patients. Researchers are working on developing cortical implants (among other types of implants), which involve the insertion of arrays of electrodes into the visual cortex. The electrodes directly stimulate the neurons, in turn producing visual percepts called phosphenes [1]. These are generally described as small, discrete spots of light appearing at predictable positions in the visual field, clearer at the centre and blurrier towards the periphery [1]. Consequently, the electrical stimulation of the visual cortex (or other locations in the visual pathway) generates 2D patterns of on-off phosphenes which, although based on a simple phenomenon, has the potential to create complex scenes.

Early research in the field focused on providing blind patients with minimal functionality such as reading text and improving their independent mobility. Throughout the following years, researchers have improved various aspects of the system, including physiological characteristics (transitioning from surface stimulation to intracortical microstimulation) and portability, with the goal of developing a “useful” artificial vision prosthesis [2].

In response to the significant neurophysiological, as well as engineering challenges posed by implants [3], researchers started simulating prosthetic vision in sighted individuals. Generally, this involves the use of an external camera and an image processing unit while the (healthy) recipient wears a head-mounted display presenting the processed image [4]. This has facilitated the research in the field and allowed investigating different aspects of phosphene vision without the associated risks posed by implants. Moreover, the freedom of adjusting and experimenting with the parameters of the models is of critical importance for developing and assessing new image processing techniques.

At early stages the image processing would consist of converting the camera input to greyscale, downsampling the result and then thresholding into binary or applying a simple edge detection filter [5]. Edge detection relies mainly on luminance, or the difference in contrast, hence much of the information present in the image is ignored. The simplicity of these methods has, in turn, an expected advantage: fast computation time. This represents an important feature of a usable visual prosthesis.

More recently, researchers started using more advanced strategies for extracting important features from the scene [6] such as saliency and depth information [4] or importance and region-of-interest maps [7]. The interest in using more complex image processing techniques originates from the fact that cortical implants have limited implantation sites [8] and long-term use leads to electrode dropout [9]. Consequently, the phosphene image perceived by the implant receiver will likely have low-resolution. Advanced processing techniques can help optimize the image quality by making use of high-level information.

One such strategy is semantic segmentation, or semantic labeling, an image processing technique used for scene understanding. In this method a neural network performs pixelwise prediction or, simply put, labels each pixel of an image with a semantic category. For this task, the fully convolutional networks (FCN) have been proposed as one of the most successful deep learning methods currently [10]. Introduced by Long et al. [11], FCNs transform the fully connected layers of a convolutional neural network into convolutional layers, hence producing spatial maps [11]. To obtain the output, the maps are then upsampled using an operation called deconvolution (see fig. 1). The architecture of this network allows for efficient training and fast inference, while at the same time it has the ability to process inputs of arbitrary size.

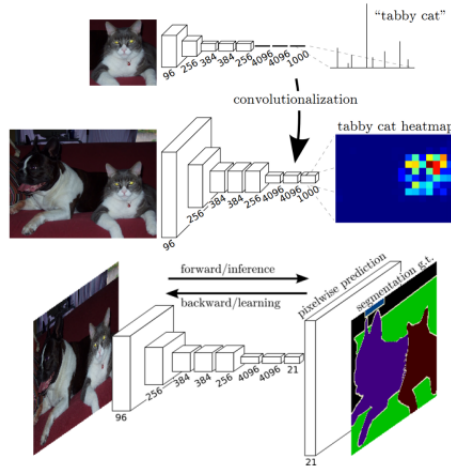


Figure 1: Fully convolutional network. Transforming fully connected layers into fully convolutional to produce heatmaps. Upsampling through the deconvolution layer allows for pixelwise labeling. Figure retrieved from [11].

Although efficient for single frames, FCNs are not fast enough for real-time inference, an important requirement for a usable prosthesis. Furthermore, they do not make use of important information available in video such as temporal continuity. The clockwork FCN proposed by Shelhamer et al. [12] is an adaptation of the FCN network: it trades a small percentage of accuracy for a big gain in efficiency. The clockwork FCN takes advantage of the fact that, for video sequences, the deeper layers of the network have a slower rate of change than shallower ones. This is based on the insight that deep layers learn global cues, while shallower layers represent more local information [12]. Layers are grouped together into stages and updated depending on a fixed or adaptive clock rate such that parameters from deeper layers can be reused for subsequent frames to save execution time.

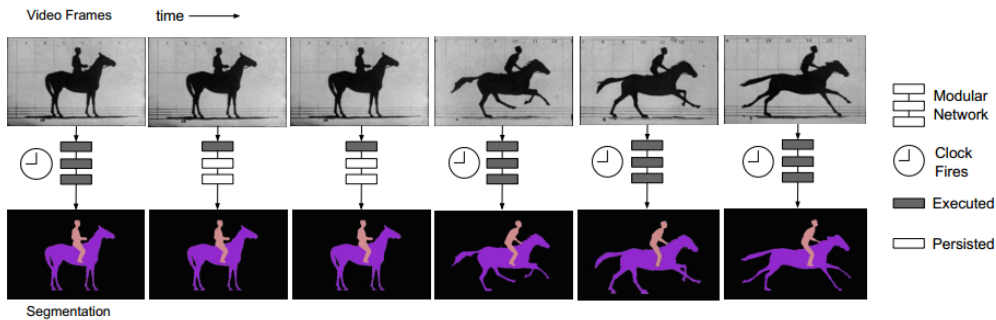


Figure 2: Clockwork method for computing frames. An adaptive or fixed clock triggers computation of frames. For static frames only the shallowest layers are computed, while the rest are persisted from previous frames. Figure retrieved from [12].

There exists evidence that more complex image processing algorithms are helpful for various tasks. For instance, Wang et al. [13] compared two complex processing methods - based on background subtraction and foreground enhancement - with a simple method that directly maps pixels to lower resolution output. The results showed improved recognition accuracy for the two complex algorithms for mobility and navigation tasks.

Parikh et al. [14] developed an efficient model based on saliency information which improved a previous model speed by ten times. The algorithm, however, obtained an execution speed of only 1.2 frames per second (fps). As underlined by Li [15], a prosthesis should be able to process the input at a frame rate of around 10Hz for the user to benefit from head movements.

A trade-off seems to exist between optimizing the representation of the physical world and the computational time required by the image processing algorithm. Given this trade-off, an important question becomes apparent: in the context of prosthetic vision, are there any benefits of using complex image processing algorithms over simpler ones? Do the advantages outweigh the disadvantages? More specifically,

*how does an advanced image processing model, which uses semantic segmentation, influence the objective performance and subjective experience of healthy subjects in a task under simulated prosthetic vision?*

To this end, we will compare a simple algorithm - represented by edge detection - with a more complex one - represented by semantic segmentation and edge detection - by means of a visual search task.

It is important to note a few focus points that help understand the motivation behind some of the choices made in the process:

1. *indoor scenes.* The well-being of a visually impaired or blind person is directly linked to self care, mobility and safety skills; her quality of life is also affected by vocational activities and daily living activities [16]. It only seemed natural to steer the focus to more basic and functional capabilities such as being able to perceive and interact with the objects in your own room.
2. *realistic.* Lab conditions under which image processing algorithms are evaluated do not always reflect real-life conditions. For instance, in [17] a lifting system was used because the headset weighted 1.5 kg; it is likely that a participant will not have the same freedom of movement using this system as when moving without it. For the same reason this research uses video as input for subjects instead of still images in an attempt to simulate real-life usage conditions of prosthesis recipients.
3. *real-time.* As noted previously, the image processing should be fast enough for the prosthesis to be usable. Here we consider 10Hz (or 10 fps) to be a good representation of a (near) real-time application.

The remaining sections of this document are structured as follows: the Methods section describes the pipeline used for generating the phosphene output, as well as the experiment carried out in order to answer the research question. Further, the results are presented and interpreted, after which a discussion follows. At last, the Future research section outlines possible improvements for the current setup and ideas for extending the current project.

## 2 Methods

### 2.1 Semantic segmentation model

#### 2.1.1 Base network

The fully convolutional network <sup>1</sup> with a stride of 8 pixels (FCN-8s) was used as the basis for the semantic segmentation model. The 8 stride version of the network was chosen in particular because it had the best pixel accuracy among all other versions [11]. Along

---

<sup>1</sup><https://github.com/shelhamer/fcn.berkeleyvision.org>

with code for training the networks, the authors also made available the learned weights obtained after the network was trained on various datasets. The datasets, however, contained mainly outdoor data. Since the focus of this project is on indoor scenes, the network had to be fine-tuned on indoor data.

### 2.1.2 Indoor data and augmentation

The ADE20K dataset<sup>2</sup> was used for this purpose. The dataset contains over 20,000 training images (outdoor and indoor), with 150 object categories, and their corresponding segmentation masks. After manually selecting the indoor scenes, the resulting dataset had little under 7,000 images.

The indoor data was augmented to approximately 300,000 images. The images were translated by 16 pixels for each new frame (i.e. the first augmented image is translated 16 pixels from the original, the second 32, etc.) in the direction of the bigger dimension, resulting in 15 new frames (or less, depending on width and height of the original image). Additionally, 11 frames were made by rotating to the left or right and back again by two degrees for every frame, up to 12 degrees. Here, the way data was augmented can also be seen as a rough simulation of the way a subject wearing a visual prosthesis would behave: translation corresponds to the subject moving to the side, while keeping her head still and rotation simulates head movements. Images were cropped to a resolution of 500 x 500 before they were fed into the network.

### 2.1.3 Finetuning

Finetuning is the process of taking an already trained model and retraining it to perform a similar task. For this project, I used the weights<sup>3</sup> of an FCN-8s network previously trained on the 2011 Pascal VOC dataset<sup>4</sup>. This dataset only contains 20 object classes.

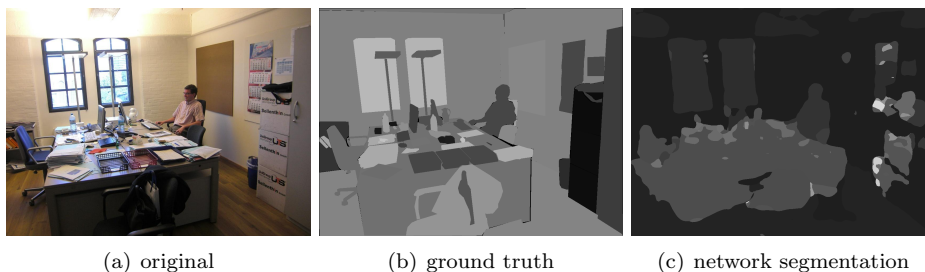


Figure 3: Example of semantic segmentation output on an image from the ADE20K training dataset. The original image and the ground truth are included for reference.

We finetuned the network using stochastic gradient descent, a batch size of 1, high momentum (0.99) and a small learning rate ( $10^{-14}$ ). As per Long et al. [11], “heavy” online learning (batch of one) with high momentum results in more accurate models in less training time. The small learning rate should prevent big “jumps” in weight updates when the network encounters an unusual image. We trained the network using the Caffe<sup>5</sup> deep learning framework.

<sup>2</sup><http://groups.csail.mit.edu/vision/datasets/ADE20K/>

<sup>3</sup>The file containing the weights (i.e. caffemodel) can be found here: <https://github.com/shelhamer/fcn.berkeleyvision.org/blob/master/voc-fcn8s/caffemodel-url>

<sup>4</sup><http://host.robots.ox.ac.uk/pascal/VOC/voc2011/index.html>

<sup>5</sup><http://caffe.berkeleyvision.org/>

Under this setup the finetuning failed, with the last layer (i.e. the softmax loss) having all weights zeroes. The forward pass to obtain the output thus resulted in nothing more than a black image. Due to time restrictions, I had to find a different solution and make compromises. I came across a caffemodel<sup>6</sup> (i.e. network weights) for an FCN finetuned on the whole ADE20K dataset; this was not ideal since finetuning on even more outdoor data likely made the network more “responsive” to outdoor scenes than indoor. Despite the drawbacks, the weights of this network were used to create the semantic segmentation stimuli for the experiment.

## 2.2 Experiment

In this experiment we investigated the performance of healthy subjects in a visual search task in simulated phosphene vision.

### 2.2.1 Participants

Sixteen participants (ten male and five female), all students of the Radboud University, with ages between 19 and 25, volunteered in this experiment. All participants had normal or corrected-to-normal vision. Subjects gave their written informed consent prior to the start of the experiment. Participants were naive with regards to the purpose of the study. Before the experiment began, they were briefed on what to expect and what their task is. They also received detailed written instructions at the beginning of the experiment and were told they can ask questions at any moment.

### 2.2.2 Setup and task

Subjects were seated in front of a laptop on which the experiment was running. The laptop’s display had a resolution of 1366 x 768. A wireless mouse was placed next to the laptop and participants were instructed to keep their hand on it at all times during the experiment, to ensure there is no delay in reaction time from external factors. With the other hand they had to press a key on the keyboard to navigate through instructions and through the experiment itself. The experimenter instructed the participants to restrict their movement during the experiment, as head movements have been demonstrated to increase performance of healthy subjects in prosthetic vision simulations [18]. The experiment was built using the PsychoPy<sup>7</sup> framework.

Participants were shown fifteen phosphene videos of indoor scenes. Each video had a length of ten seconds. Before each video, a target word representing a common household object was shown on the screen. The subject’s task was to find the target object in the video and click on it as fast as possible and as accurately as possible. They were encouraged to take a guess before the video ended even when they were not confident with their choice. After each video, participants were asked to rate the confidence of their choice by means of a slider with a scale between 0 and 100.

Prior to the experiment trials participants had a chance to practice on three videos. After they clicked on the video, the video stopped and the phosphene frame at which the video was clicked was displayed on the screen, as well as the original frame, shown side by side. No explicit feedback (i.e. whether they clicked on the target object or not) was given. Participants were free to inspect the original frame for as long as they wanted.

The experiment follows a between-subject design, with eight subjects for each of the two following conditions:

---

<sup>6</sup><http://sceneparsing.csail.mit.edu/model/caffe/>

<sup>7</sup><https://www.psychopy.org/>

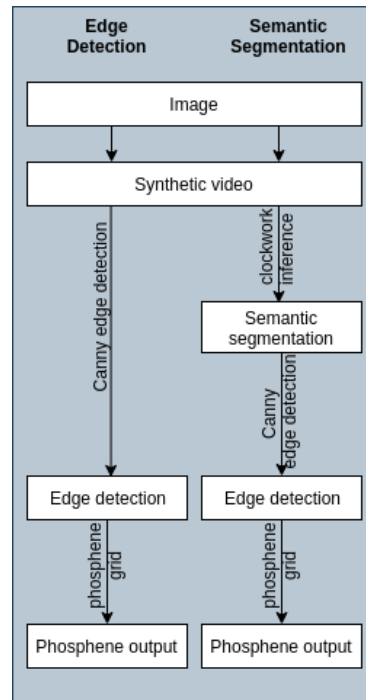
1. *edge detection*
2. *semantic segmentation*.

Before the experiment began, each participant was randomly assigned to one condition; however, some have been assigned to a certain condition such that the balance of subjects per condition was maintained.

### 2.2.3 Stimuli

In this section we will present the procedure used to obtain the stimuli for the experiment. Figure 4 gives an overview of the pipeline.

Figure 4: Pipeline for obtaining the experiment stimuli for the two conditions. For both edge detection and semantic segmentation the first step is identical; after synthetic videos are obtained, the semantic segmentation track has an extra step, that is obtaining semantic segmentation using clockwork inference. The Canny edge detection algorithm is then applied to both intermediate results. Lastly, the phosphene outputs representing the stimuli are obtained for each condition by multiplying with a phosphene grid.



The pipeline starts with images of indoor scenes acquired from Flickr and Pexels<sup>8</sup>. Figure 5 shows some of these images. They represent real-life indoor scenes with no lighting modifications and are meant to simulate the natural environment in which an artificial prosthesis would be used. In addition, we also used four stock images (characterised by minimalistic scenarios with few objects in the frame and unnatural lighting settings) for comparison purposes. All target objects have been “seen” by the semantic segmentation network, i.e. the network was trained on images which contained the target.

<sup>8</sup><https://www.flickr.com/> and <https://www.pexels.com/>



Figure 5: Examples of real-life (top left, top right, bottom left) and stock (bottom right) images. Here, the target objects are chair, mug, table and lamp, respectively.

From each image we then manually generated<sup>9</sup> synthetic videos by first translating and then zooming in using the Adobe AfterEffects<sup>10</sup> software. The synthetic videos have 100 frames and a duration of 10 seconds (each), thus resulting in a 10 frames per second (fps) rate. The resolution was fixed to 600 x 375. As mentioned previously, a frame rate of 10 fps or more is an important characteristic of a real-time prosthetic vision application. Using the same procedure, masks were generated for the target object in each video; these have been used to assess the correctness of subjects responses.

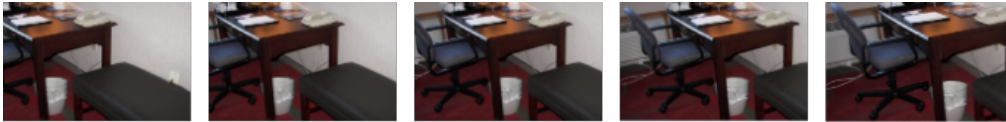


Figure 6: Synthetic video generation using translation and zooming in. Only frames 1, 25, 50, 75 and 100 are shown here.

Following the pipeline, the next step for generating stimuli for semantic segmentation condition is performing the actual segmentation on our synthetic videos. The inference (i.e. obtaining the output) uses a clockwork strategy which can be summarized as follows:

- for one frame: forward pass in the usual way (that is, feed the network the frame and get the output)
- for the next frame: compute output for the shallowest layers and merge it with the output for the remaining layers from the previous frame
- repeat for all remaining frames.

This way the network only computes a full forward pass for half the number of frames and reuses useful information from previous frames to reduce computation.

We can now apply edge detection to the synthetic videos and to the semantic segmentation network output. An example of edge detected frames can be seen in the figure below.

<sup>9</sup>With enormous help from Constantin Börker and his AfterEffects skills

<sup>10</sup><https://www.adobe.com/products/aftereffects.html>

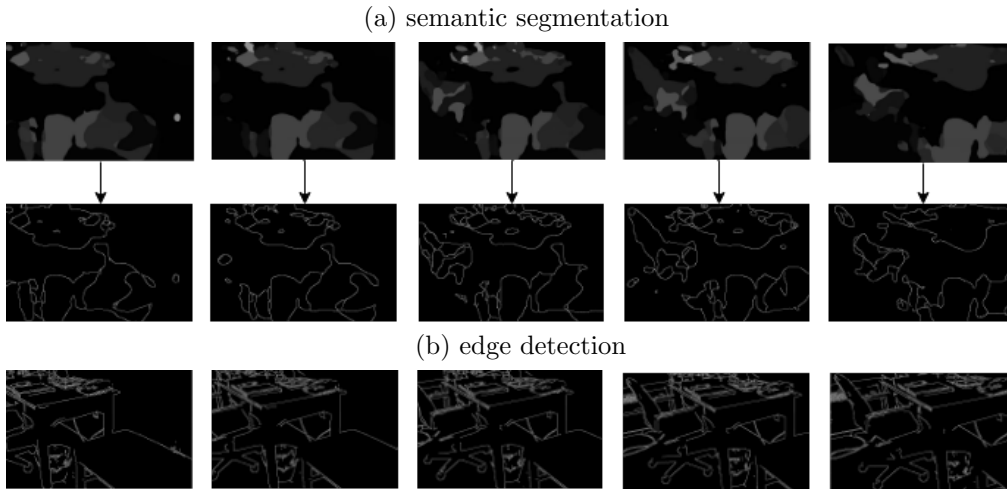


Figure 7: Example output for both conditions after applying edge detection. These are the same frames as the ones used in Figure 6.

- (a) first row: output semantic segmentation model
- second row: edge detection output
- (b) edge detection applied directly to synthetic videos

We used OpenCV's Canny edge detection algorithm<sup>11</sup>.

The final step of the pipeline is obtaining the phosphene output. We compute it by multiplying the edge detected videos described above with a grid of phosphenes (shown in fig. 8). The code for the grid was written by Caroline Bollen and refined by Joeri Hartjes. We further refined the code to obtain a grid of 40 x 25 (1000) phosphenes. The resulting image has a resolution of 80 x 50.



Figure 8: Phosphene grid

The phosphenes in the grid vary in size, in accordance with implant recipients observations [1]. This creates a clearer grid center and a blurrier periphery.

The phosphene output, as presented in the experiment, has a resolution of 160 x 112.

<sup>11</sup>[https://docs.opencv.org/3.1.0/da/d22/tutorial\\_py\\_canny.html](https://docs.opencv.org/3.1.0/da/d22/tutorial_py_canny.html)

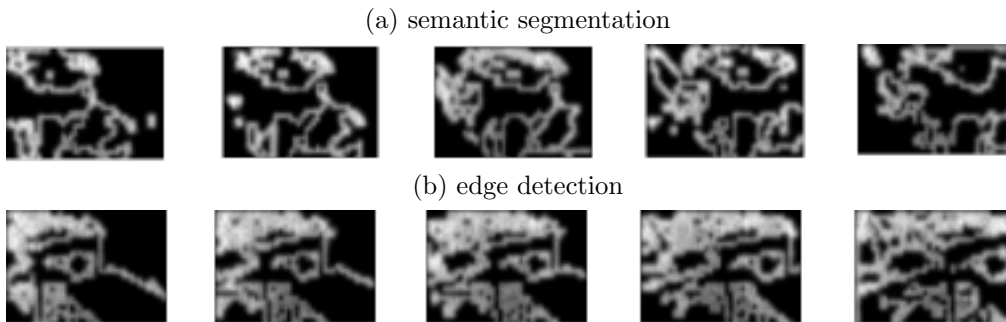


Figure 9: Final phosphene output (experiment stimuli) for each condition.

### 3 Results

The reaction time for correct responses is plotted in the boxplots from figure 10. The edge detection group has a mean of 3.15 and a median of 2.59 seconds, while the semantic segmentation group a mean and median of 5.96 seconds. The edge detection condition contains a few outliers, with reaction times almost two seconds slower than the non-outlier responses from the same condition. Since the graph shows individual responses (i.e responses from every subject for every video stimulus), it is possible that one certain target object was particularly hard to detect in a video and so the responses for this stimulus might represent the outliers. We are also interested in results from these responses, therefore we did not remove these outliers.

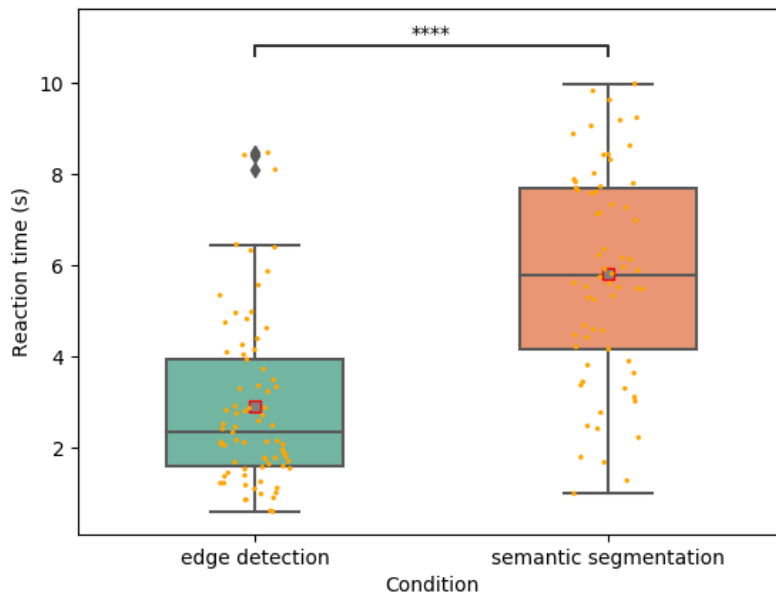


Figure 10: Reaction time boxplots for correct responses, grouped by condition.

The semantic segmentation boxplot shows a large variance from 1 second to 10 sec-

onds, with responses evenly distributed between these two values. The boxplot for edge detection condition shows a smaller distribution; the best response for edge detection is also faster than the one in the semantic segmentation group. The result of an independent t-test is shown on the plot as well. The four stars indicate a  $P$  value smaller than .001, while the  $t$ -value is -8.38.

Figure 11 shows the accuracy for each subject per condition. Here we see one outlier with an accuracy of 0.28 in the edge detection condition. The maximum accuracy for both conditions reaches 0.8; however, there is more variance for semantic segmentation than for edge detection, with two subjects having an accuracy below 0.4. The means for the two conditions is surprisingly close, which is indicated by the not significant result of the independent t-test ( $p$ -value = 0.59 and  $t$ -value = 0.54). If we disregard the outlier in the edge detection condition, the accuracy for subjects in that group are, remarkably, all above or equal to 60%.

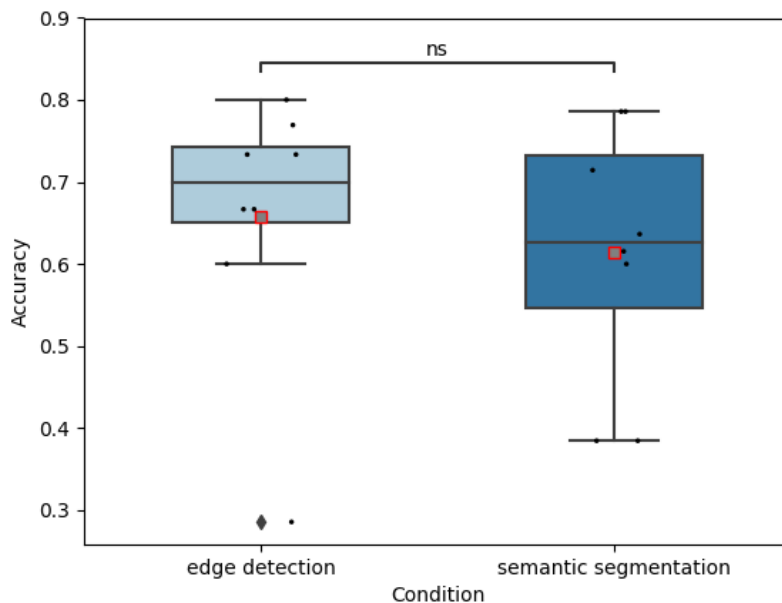


Figure 11: Accuracy for each participant, grouped by condition.

Subjects' confidence in their choice was also measured. This is plotted in fig. 12, which only shows the confidence for correct responses. The plot for confidence for all responses showed a similar trend as the one displayed here and therefore it was omitted.

Participants for both conditions show much variance in confidence, with the edge detection subjects having a slightly bigger median (and mean). An interesting aspect here is that half of the responses in the semantic segmentation group are below 50%. The independent t-test gives a  $p$ -value smaller than .001 and a  $t$ -value of 6.11.

Reaction times for correct responses for real-life and stock stimuli are also shown in fig. 13. As expected, for edge detection the mean and median response times are slightly better for stock than for real-life stimuli, although the distributions are almost identical. Responses in semantic segmentation condition show, surprisingly, a wider distribution, with two responses for stock stimuli having a bigger reaction time than the worst reaction time for real-life stimuli. The mean and median values for semantic segmentation are similar for the two types of stimuli.

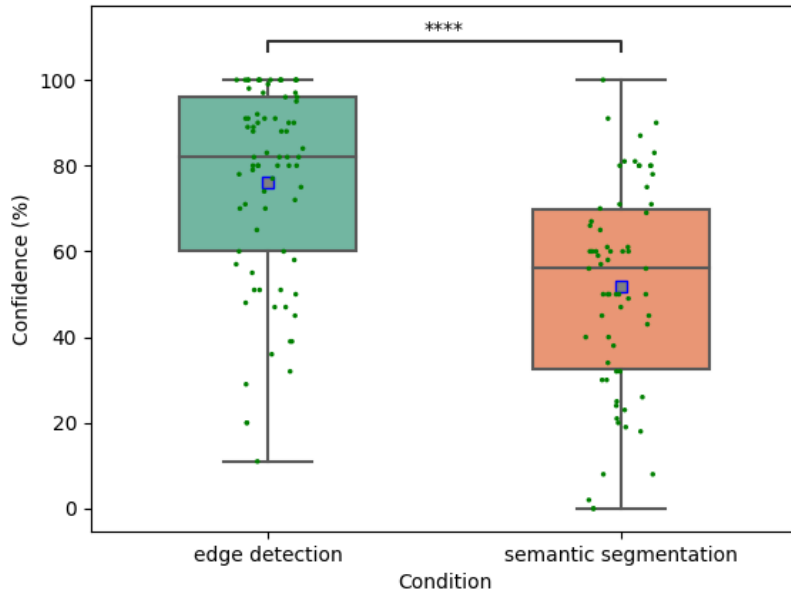


Figure 12: Confidence for correct responses, grouped by condition.

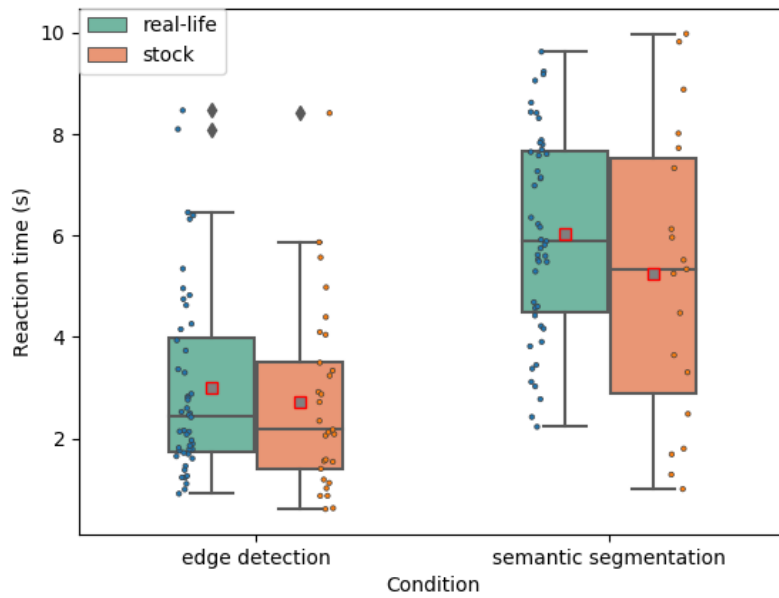


Figure 13: Reaction time boxplots for correct responses for complex (i.e. *real-life*) vs simple (i.e. *stock*) scenes, grouped by condition.

## 4 Discussion

In the beginning of this paper we raised the question whether it is advantageous to use complex image processing algorithms instead of simple ones for processing the input of a prosthetic vision system. We compared a simple algorithm, edge detection, with a more complex one which uses semantic segmentation in a visual search experiment on healthy subjects. We were interested in both their objective performance (e.g. accuracy and reaction time), as well as their subjective experience during the experiment. Before we interpret the objective results we will discuss the observations gathered by the experimenter during and after the experiment.

The subjects were asked, after the experiment, how they perceived the task. The majority of subjects thought the task was difficult, but not tiring or frustrating (with one notable exception, one participant who asserted that the task was easy and he was always confident in his answer). There were, however, also subjects who verbally expressed dissatisfaction because they could not find the target within the ten available seconds. Most of these subjects were in the semantic segmentation condition. The subjects from this condition also seemed to be taken by surprise the most by the task in the first trials. Conversely, there were also subjects in the edge detection condition who thought the task was very difficult. Generally speaking, participants perceived the task in a wide variety of ways, with a larger number of participants in the semantic segmentation group showing more negative emotions while performing the task. This is also reflected in the lower median confidence expressed by the semantic segmentation group than the edge detection one (see fig. 12).

Participants were also asked if the practice trials prior to the experiment helped with the experiment task. A large number of subjects thought the practice trials indeed helped them picture the scenes more abstractly. This, especially for subjects in the semantic segmentation condition, also made it easier for them to overcome some inaccurate representations of the target objects in phosphene vision. Figure 14 shows such an example where the semantic segmentation model segments the lamp in two pieces; a subject looking for a lamp will likely be confused at first by this representation. With practice, however, the subject learns to “fill in” the gaps in the phosphene representation of the object. This capacity for abstraction might, in fact, also explain the small difference in accuracy between the two conditions (fig. 11). Nevertheless, this comes at the price of slower reaction times (fig. 10).

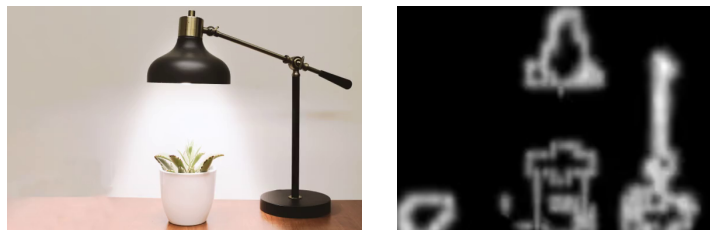


Figure 14: Example of inaccurate representation of target object in stimulus for semantic segmentation. The left image shows the first frame of the original video and the right one shows the phosphene output for the semantic segmentation condition (for the same frame).

The arm of the lamp in the phosphene frame shows discontinuities and thus it will be perceived as two separate objects.

The potential for using semantic segmentation as a tool to better understand the environment and, therefore, better represent it in phosphene vision led us to hypothesize that this image processing technique will give better results in our task, at least qualitatively. Figure 7, however, shows a much clearer, more defined result for edge detection

rather than for semantic segmentation. The poor quality of the semantic segmentation representation of the stimuli is also reflected in the number of stimuli for which subjects did not respond: out of 16 total no-responses, 13 come from the semantic segmentation group. This could be explained by a big difference in subjects performance in the groups, but is arguably more likely to be a result of indistinguishable scenes resulting from semantic segmentation, which the subject cannot even guess at.

Reaction time and confidence responses are also in line with the observations above. Reaction times for correct responses are smaller for the edge detection group than for the semantic segmentation group. While the subjects in the edge detection condition need approximately three seconds to correctly identify the target object, subjects in the semantic segmentation condition need almost twice that time.

With the current setup and model used for semantic segmentation we could not prove that using semantic segmentation has any advantages over the simpler edge detection algorithm. In reality, edge detection is a straightforward, efficient processing method which should not be overlooked or underestimated.

Although the results indicate that our semantic segmentation model did not perform to the expectations, it is for sure not the case that all other semantic segmentation models will perform worse than edge detection. Semantic segmentation is still an invaluable method for representing the environment more accurately in low resolution setups due to its capacity for scene understanding. Our conclusion is certainly not definitive and should be taken with caution, especially considering the limitations presented in the following section.

## 4.1 Limitations

The main limitation of our study is the semantic segmentation model itself. More specifically, the dataset used to finetune the model contained too little indoor data. The model performed only as good as the data it was trained on. If our focus would have been on outdoor data, then the model would perhaps perform better. Unfortunately, we had to use this model since we could not get results from our finetuned model. Additionally, the number of participants in the experiment would need to be higher in order to obtain statistically significant results.

Furthermore, measuring inference times to obtain the stimuli for both edge detection and semantic segmentation would have given a broader view on the advantages and disadvantages of these methods. As mentioned in previous sections, the computational time needed to infer the stimuli is of great importance for a usable prosthetic vision system. Since we did not measure the inference time for semantic segmentation stimuli, it is possible that the network requires longer processing times than our requirement of 10 frames per second, making this approach unusable for real-life applications.

Lastly, although we clearly state that we strive for our experiment to be as realistic as possible, there are a few details in our setup that do not fulfill this requirement. For instance, participants do not have much freedom of movement in our experiment. In a real-life situation, a patient wearing a visual prosthesis is free to explore the environment and use head movements to gather more information, as proved to enhance performance [18]. On the same note, the wearer would likely be aware of the dimensional perspective. If, for instance, she is standing and looking down towards a desk, the target objects will be represented differently than if she would be sitting. This information would presumably help the patient distinguish objects easier. In our experiment the subjects do not have this information and changing perspectives (e.g. fig. 16) can be confusing. Nevertheless, the last discussed aspect does not influence our results since both experimental groups faced the same challenge.

## 5 Future work

The results from our experiment are likely a consequence of the poor semantic segmentation model used for generating the stimuli. Consequently, future research groups could focus on finetuning the fully convolutional network as described in this paper, on appropriate data (i.e. indoor data if the focus is on indoor applications).

The clockwork inference architecture would also be a potential improvement point. The authors of the clockwork network proposed a fixed clockwork schedule (such as the one we used in this research), as well as an adaptive one [12]. The adaptive clockwork would further make use of temporal continuity in video frames and change the update rates accordingly. This would increase the network accuracy, although for very dynamic frames the computation would become heavier.

Moreover, the experiment could be performed using a more realistic environment, such as virtual reality (VR). VR is still the most common instrument for simulating the experience of a real prosthesis in healthy subjects [6].

## 6 Conclusion

Using a finetuned semantic segmentation network based on fully convolutional networks and a clockwork architecture for inference, we could not prove that an image processing method based on semantic segmentation is superior to the simpler edge detection algorithm. Edge detection remains an efficient, robust and straightforward technique for image processing which yields clear and reliable results.

## References

- [1] G. S. Brindley and W. S. Lewin, "The sensations produced by electrical stimulation of the visual cortex," *The Journal of Physiology*, vol. 196, no. 2, p. 479–493, 1968.
- [2] W. H. Dobbie, "Artificial vision for the blind by connecting a television camera to the visual cortex," *ASAIO Journal*, vol. 46, no. 1, p. 3–9, 2000.
- [3] P. M. Lewis, H. M. Ackland, A. J. Lowery, and J. V. Rosenfeld, "Restoration of vision in blind individuals using bionic devices: A review with a focus on cortical visual prostheses," *Brain Research*, vol. 1595, p. 51–73, 2015.
- [4] W. L. D. Lui, D. Browne, L. Kleeman, T. Drummond, and W. H. Li, "Transformative reality: Augmented reality for visual prostheses," *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011.
- [5] L.-X. Buffoni, J. Coulombe, and M. Sawan, "Image processing strategies dedicated to visual cortical stimulators: A survey," *Artificial Organs*, vol. 29, no. 8, p. 658–664, 2005.
- [6] 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.
- [7] J. R. Boyle, "Region-of-interest processing for electronic visual prostheses," *Journal of Electronic Imaging*, vol. 17, no. 1, p. 013002, 2008.
- [8] N. Srivastava, P. R. Troyk, V. L. Towle, D. Curry, E. Schmidt, C. Kufta, and G. Dagnelie, "Estimating phosphene maps for psychophysical experiments used in testing a cortical visual prosthesis device," *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, 2007.
- [9] M. S. Humayun, J. D. Dorn, L. da Cruz, G. Dagnelie, J.-A. Sahel, P. E. Stanga, A. V. Cideciyan, J. L. Duncan, D. Elliott, E. Filley, A. C. Ho, A. Santos, A. B. Safran, A. Arditi, L. V. D. Priore, and R. J. Greenberg, "Interim results from the international trial of second sights visual prosthesis," *Ophthalmology*, vol. 119, no. 4, pp. 779–788, 2012.
- [10] A. Garcia-Garcia, S. Orts, S. Oprea, V. Villena Martinez, and J. Rodríguez, "A review on deep learning techniques applied to semantic segmentation," 2017.
- [11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [12] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," *Computer Vision–ECCV 2016 Workshops*, 2016.
- [13] J. Wang, Y. Lu, L. Gu, C. Zhou, and X. Chai, "Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies," *Information Sciences*, vol. 277, pp. 512–524, 2014.
- [14] N. Parikh, L. Itti, and J. Weiland, "Saliency-based image processing for retinal prostheses," *Journal of Neural Engineering*, vol. 7, no. 1, p. 016006, 2010.
- [15] W. H. Li, "Wearable computer vision systems for a cortical visual prosthesis," *2013 IEEE International Conference on Computer Vision Workshops*, 2013.

- [16] G. Dagnelie, "Psychophysical evaluation for visual prosthesis," *Annual Review of Biomedical Engineering*, vol. 10, no. 1, p. 339–368, 2008.
- [17] M. J.-M. Macé, V. Guivarch, G. Denis, and C. Jouffrais, "Simulated prosthetic vision: The benefits of computer-based object recognition and localization," *Artificial Organs*, vol. 39, no. 7, 2015.
- [18] S. C. Chen, L. E. Hallum, G. J. Suaning, and N. H. Lovell, "Psychophysics of prosthetic vision: I. visual scanning and visual acuity," *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006.

## 7 Appendix

Target object														
notebook	mouse	pillows	fruits	laptop (stock)	chair	sofa	mug (stock)	pillows (stock)	flowers	table	nightstand	lamp	keyboard (stock)	laptop

Table 1: Order of stimuli used in the experiment. In both conditions the order remained the same, as well as between subjects

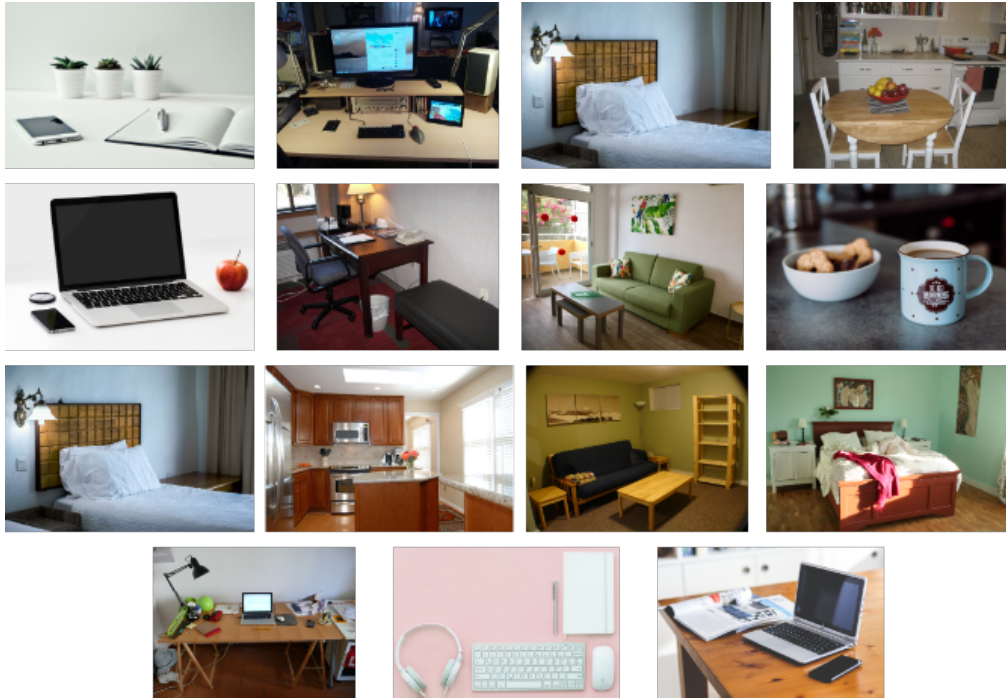


Figure 15: Original images used as basis for stimuli. They are arranged in the same order as the table above.

Target object		
lamp (stock)	mug	keyboard

Table 2: Order of stimuli used for practice trials. In both conditions the order remained the same, as well as between subjects



Figure 16: Original images used for practice trials. They are arranged in the same order as the table above.

*Note:* you can find the code for training the semantic segmentation model, data augmentation, as well as the stimuli for the 2 conditions, experiment data and analysis in the repository <https://github.com/codrutalugoj/Bachelor-Thesis-Phosphene-Vision-and-AI>.