

# Optical Prosthesis Image Processing Using Computer Vision and Convolutional Neural Network

**Asmi Choudhary**

Class of 2023, Delhi public School Kuwait, 49 South street, Ahmadi, Kuwait

Email-id: [asmichoudhary03@gmail.com](mailto:asmichoudhary03@gmail.com)

**Vinay Vishwakarma**

Research and Innovation, Roboscience Education Labs Pvt. Ltd., Lokhandwala, Oshiwara, Mumbai, India.

Email-id: [vinay.vishwakarma@omotec.in](mailto:vinay.vishwakarma@omotec.in)

## **Abstract:**

Optical prosthesis is a way to restore vision to millions of people who lost their eye-sight due to diseases or accident causing degradation of vision. The optical prosthesis device transforms the recorded images into corresponding electrical stimulation patterns, which are then used to create phosphenes. However due to some uncertainty in the internal electrodes the induced perception is far from ideal. Therefore, in this study a novel approach is proposed that can convert the object from video feed into phosphene image. The proposed approach comprises of four phases. The proposed approach extracts frame by frame of the video feed and recognizes the object images with the help of a pre-trained mask-RCNN model. The objects identified in the images are separated from the background by semantic segmentation. Then the object images are converted into phosphene images which are then superimposed to recreate the scene. The proposed approach is repeated for each frame of the video. The strength of a proposed model lies in its practical applicability. Therefore the approach is experimentally run on a video and tested. The result obtained from the experimentation can confirm that the proposed model is effective as well as efficient.

**Keywords:** Computer vision, Convolutional Neural Network, Retinal Prosthesis, Mask-RCNN,

## **1. Introduction**

A prosthetic device called a retinal implant converts visual pictures into control signals and then stimulates the remaining retinal circuitry using those signals. Image compression for bionic eyes shrinks and resizes the images while maintaining their object detection rate. The generated scaled image's object detection rate is comparable to the rate of the original image. The processing overhead for implants inside the body is indeed reduced by this. The 130 million photoreceptors in the retina of the eye are converted into electrical impulses and transported by 1.2 million highly specialized ganglion neurons, whose axons make up the optic nerve. The main visual cortex of the brain receives visual information from the optic nerve via the lateral geniculate nucleus. The device can be placed in the 1) retina, 2) visual cortex and 3) optic nerve. The retina is further broken down into the sub-retinal and Epi-Retinal layers [1]. The sub-retinal refers to the side that is next to the choroid, and the Epi-Retinal refers to the side that faces the vitreous. The visual cortex, the optic nerve, and three. Due to its many benefits, Epi-Retinal Stimulator is the subject of most research institutes efforts.

The major drawback in the commercially available Epi-Retinal Implant (ERI) costs about USD 100 thousand fails to restore the necessary vision due to its 60 electrodes resolution [2]. In the present state-of-the-art literatures retinal prosthesis is mostly concentrated on developing implant with hundreds of electrodes for the blind to read letters, navigate a room and recognize faces [3 – 5].

### **1.1. Motivation and Novelties**

Motivated by the absence of a cost effective technical approach to enhance the visual perception and ability to conduct visual tasks daily. This study presents an image processing technique where the captured image by the camera is compressed which is then processed for smoothing and edge detection for object and facial recognition, scene recreation and optical flow estimation. The process are conducted with the help of computer vision. However the internal implant design with real electrodes for electrical stimulation is beyond the scope of this paper.

Remainder of the paper is drafted as follows:

Section 2 reviews the contemporary literatures. Section 3 discusses the preliminary concepts used developing the proposed approach and the result obtained from the methodology. Finally section 4 is the conclusion of the paper.

### **2. Review of the Contemporary Literatures**

In the recent time, image processing along with computer vision (CV) has become a popular technique for prosthetic perception in retinal prosthesis (RP). In order to provide more helpful information for diverse visual activities and eventually enhance functional vision for the blind, CV aims to optimize prosthesis perception [6]. In CV helpful information is simply thought of as details that aid in "understanding" a setting [7]. For various applications, including facial recognition for mobile payments and industrial robot localization and obstacle detection, CV strives to automatically collect, analyze, and interpret useful information [8]. With the long-term objective of being as adept at focusing on and naturally comprehending visual information as humans, the methodologies or models developed in this field are bio-inspired by the human vision system. Some of these, like object identification, face recognition, and scene perception, have been included into simulated prosthetic systems for functional vision [9]. Saliency detection and machine learning have advanced the most among these techniques because of the way in which their respective fields of application overlap with the goal of prosthetic vision optimization. By calculating contrast from picture attributes like color, brightness, or gradient, saliency detection is a type of technique or algorithm that simulates a human's capacity to extract a region or object of interest (OOI) [10]. If the model has been trained beforehand using batches of labelled data samples, machine learning can automatically recognise objects in real time with high levels of accuracy.

### **3. Preliminaries and Methodology**

#### **3.1 Image Segmentation**

Image segmentation, also known as pixel-level classification, separates an image into groups of pixels. An image object, which is more precisely a collection of pixels, is categorized based on properties like color, depth, or intensity [11]. A number of image objects are produced as a result of this procedure, producing the grouped attributes that make up the image. The edge detection method, which is tasked with identifying the contours of objects in images based on differences from other pixels in the region, yields the outline of the image object as one property offered in these segments [12]. Edge detection and image segmentation can be useful tools for artificial vision research because they help patients understand their surroundings, including hazards like potential collisions and things like furniture [13]. The information offered by processed picture objects is used in the construction of automated robots and self-driving cars to help evaluate what information in a situation is required when taking into account the temporal limits of retinal prosthesis.

#### **3.2 Object Recognition**

It is important to pay attention to objects in images, but when using computer vision systems, the object's nature might also be very important. Some items could offer pertinent information about the surrounding environment. These elements highlight both potential dangers and the image's key features. This ability to recognize threats or

interests results in increased situational awareness, whether the knowledge is used by a person or a machine. A task like identifying an object before touching it or noting a probable collision, two capacities that were previously lacking or unattainable for the visually handicapped, can now be completed because to this improved situational awareness. Facial structures can be a complex item made up of several moving points of interest. It can be difficult to replicate this level of detail, although there have been some encouraging advancements, such as approach proposed in for low vision systems like RPs [14].

### **3.3 Scene Reconstruction**

Scene reconstruction, an aspect of computer vision (CV) systems, seeks to rebuild a scene, including specifics like depth. This depth recovery can be used for a variety of things, such helping to identify objects and accurately recreate environments [15]. However, it is not always necessary to recreate an area in great detail. In time-constrained applications, a scene can be rebuilt so that the edges of objects obtained by image segmentation reproduce a passable representation of the scene [16, 17]. The ability to discern which elements in a scene are most important remains a persistent difficulty in artificial vision.

### **3.4. Optical Flow Estimation**

Dynamic objects in an environment present various difficulties for artificial vision systems, which are already confined by time. This procedure can be aided by image segmentation, which tracks the silhouette of the item, similar to the method used in [15]. It is vital to remember that optical flow only records apparent motion and not actual item motion. Occlusions, in which one object may be covered by another, are one of the difficulties posed by dynamic objects. Such occlusions might make it difficult to distinguish between different objects in the scene, leading to poor image segmentation. When working with dynamic items, especially in unpredictable contexts like outdoor sceneries, occluding one object with another is to be expected. Optical flow algorithms may be employed to address this issue, perhaps to comprehend the depth of the scene and object layering. Understanding the depth of the environment gives the option to prioritise object recognition based on the layer it is in relation to other items; the less obscured the layer, the better the chance of recognition [17].

### **3.5. Methodologies**

Computer vision (CV) techniques are the focus of recent approaches in retinal prosthesis (RP) systems, utilizing various neural networks, usually convolutional neural networks (CNNs).

#### **3.5.1. Computer Vision**

Computer vision interprets information from media-based data, such as photographs and videos, and is used in applications like self-driving automobiles, medical diagnosis, and manufacturing. The computer can gain a better grasp of the objects or scene that it is given as input by using this data. The computer is better able to carry out automatic activities or deliver findings to an operator with a deeper knowledge. This process can be broken down into three steps 1) input of image based media, often gathered via a camera 2) processing of image data, utilizing a NN for pattern finding 3) the processor returns the findings to the requester for further actions. Unlike the human visual system, which passively gathers data, CV uses vast amounts of data to extract characteristics or patterns and frequently uses Neural Networks (NNs) as the model's foundation. This means that while computer vision applications may be very good at specific tasks like spotting flaws in parts produced on an assembly line, they may not be as good at more general tasks like understanding objects in different contexts, which is essential for patients receiving retinal prostheses.

#### **3.5.2 Convolutional Neural Networks**

Convolutional Neural Networks (CNN) belongs to the family of NNs. CNN belongs to deep learning (DL), which is often characterized as a NNs with numerous hidden layers [18]. CNN has many structures in addition to

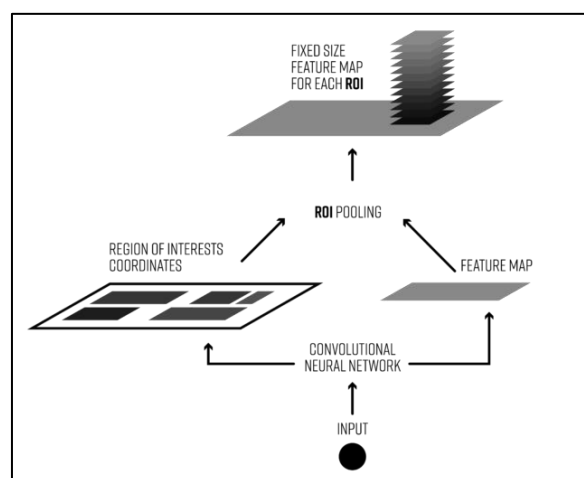
additional hidden layers. Although there may be more than one of this kind of layer, these NNs are first made up of a convolutional layer. Additionally, this is where the majority of calculation is finished and where the CNN's foundation is built. The data and a filter are the two components that are required here. We may take into account information like the height, width, color, and depth of a color image in the event of an image input. With the help of this information, we can apply a feature detector—also known as a kernel or filter—which applies a small matrix of weights to the entire image. It will then move throughout the image to check if the feature's fundamental components, such as a bright spot or a specific, are there. This is known as a convolution. The feature detector scans a little portion of the image at a time before moving on to the entire thing. Dot products, which are calculations of the input pixels and the filter, are created for each segment. When the feature detector has finished scanning the image, a feature map—a collection of dot products—is the result. In order for the model to extract patterns, the convolutional layer will ultimately transform the image into numerical values. The pooling layer, also known as down-sampling, comes after the convolutional layer. The model in this case minimizes the input parameters. The main distinction between this layer and the convolutional layer is that when the pooling process sweeps over the image, the filter has no weights. Instead, the filter performs a calculation or aggregate function on the values, which are then utilized to fill the output array. The more prevalent max pooling method and average pooling method are the two most often used methods of pooling. While average pooling determines the average of the current features, max pooling will choose the most prominent feature from the feature map. The pooling layer cleans out extra noise from the image, in contrast to the convolutional layer which only considers a portion of the image. Despite being the main cause of lost data, this activity helps CNNs by increasing productivity, simplifying problems, and lowering the danger of over-fitting the model. The fully connected layer is the top layer of these neural networks. Based on the information acquired from earlier levels, the extracted characteristics are classified in this area. The fully-connected layer often employs an activation function, such as softmax, to categories inputs and normalize outputs by putting probabilities on the sum of the weighted values.

### 3.6. Proposed Approach

The methodology proposed in the study comprises of four phases. The flow of the proposed approach is as follows:

**First phase:** Detecting of the Region of Interest (ROI).

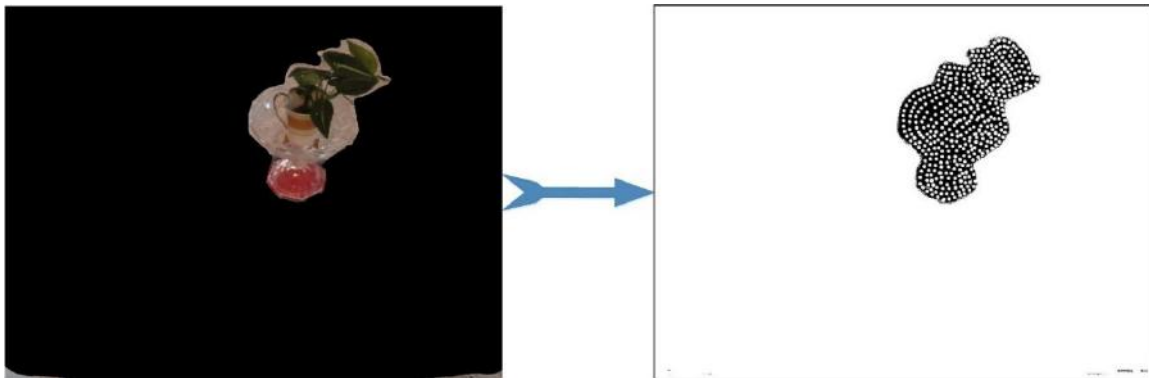
In the study, the ROI is detected by the process of Region of interest pooling (ROI pooling) [19, 20] which is an operation widely used in object detection tasks using CNN. Two major tasks in CV are object classification and object detection. In the first case the system is supposed to correctly put the bounding box around the dominant object in an image. In the second case it should provide correct labels and locations for all objects in an image. Figure 1 shows the object detection pipeline with ROI pooling.



**Figure 1:** Schematic representation of the object detection pipeline with ROI pooling

**Second phase:** Detecting edge of the object

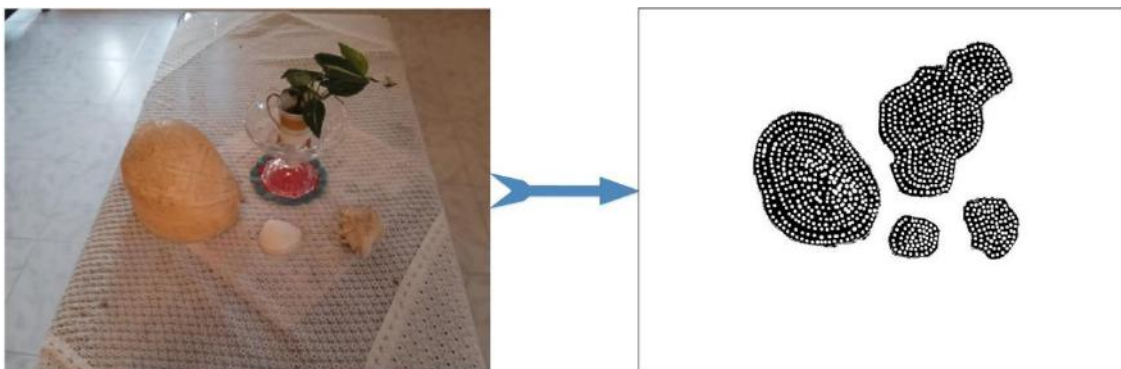
The second phase of the proposed approach is to detect the edge of the objects in the images. In this process, all the objects detected by the ROI pooling in the first phase are separated from the background by image segmentation and then the edges are detected by the edge detection in image processing techniques. Semantic segmentation is done for separating the objects from the background. Then the edge detection of the objects are conducted which is then converted to binary image. The binary image is simulated through the prosthetic vision to convert it into a phosphene image. Figure 2 represents the conversion of a binary image into a phosphene image.



**Figure 2:** Binary image into a phosphene image.

**Third phase:** Scene recreation

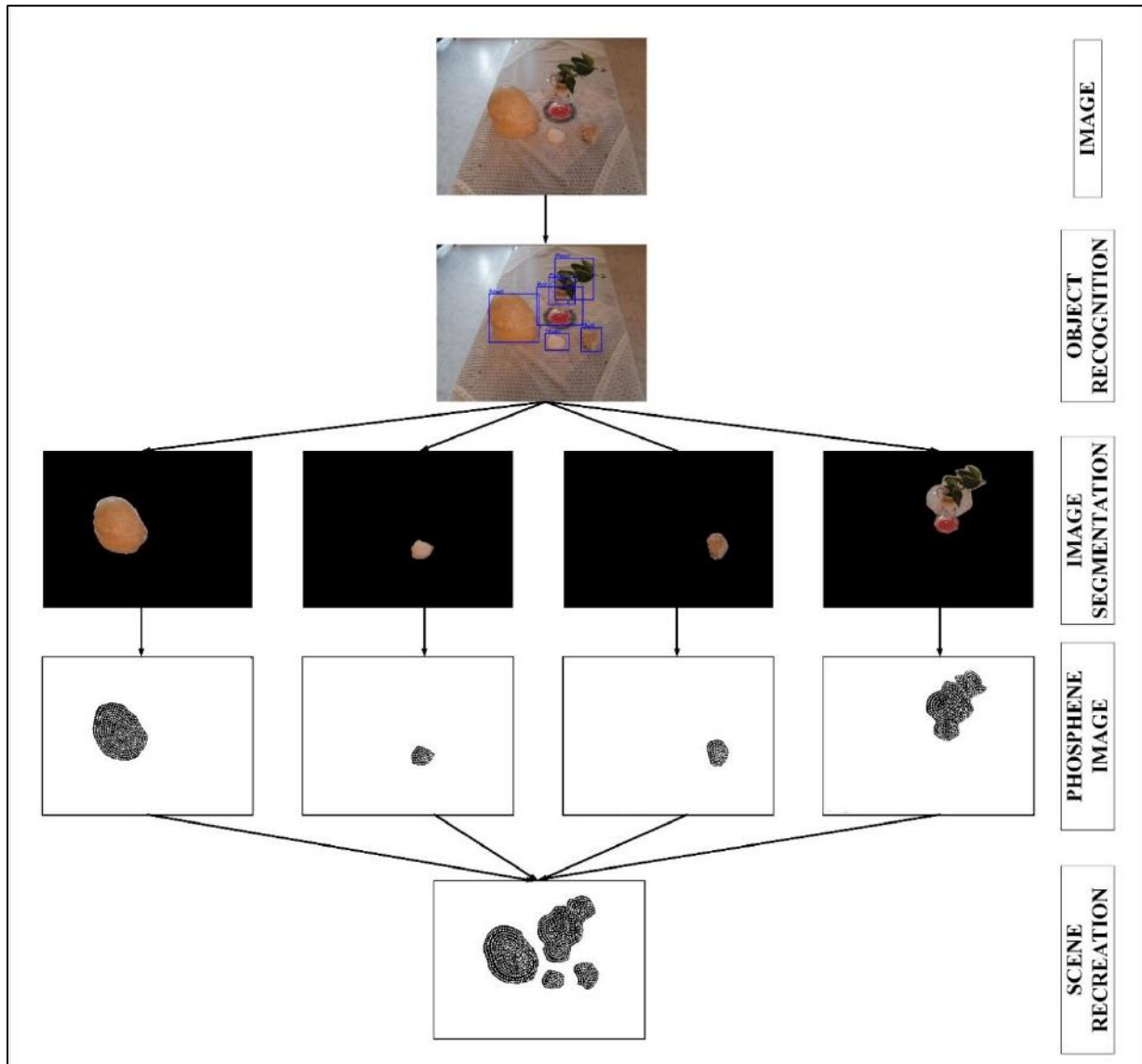
In this phase of the proposed approach all the objects that were identified in the first phase by the ROI pooling are superimposed to a single image. Figure 3 represent the scene recreation of the phosphene images of the objects.



**Figure 3:** Scene recreation of the phosphene images of the objects

**Fourth phase:** Optical flow estimation

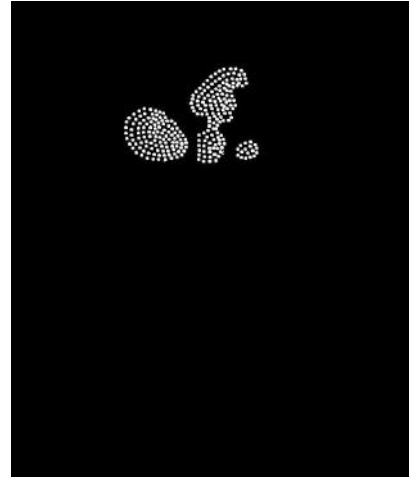
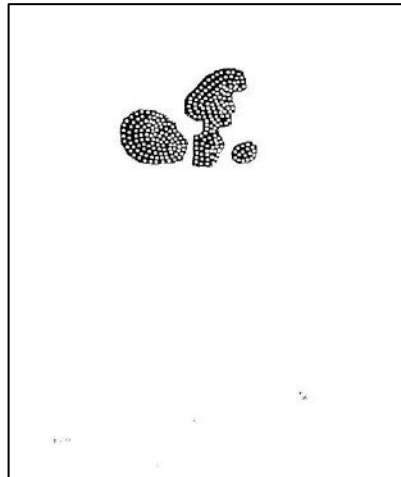
This is the final phase of the proposed approach. In this phase, the flow of the images frame by frame of the video feed is converted to phosphene images. Figure 4 shows the flowchart of the proposed approach.



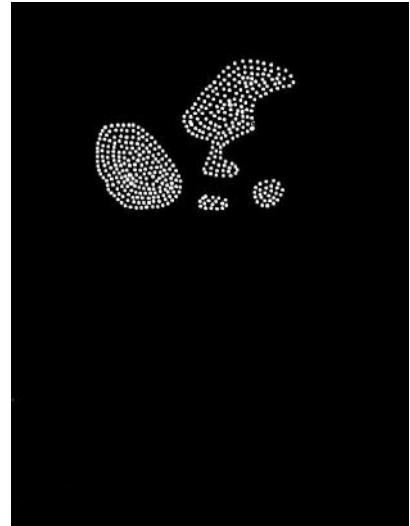
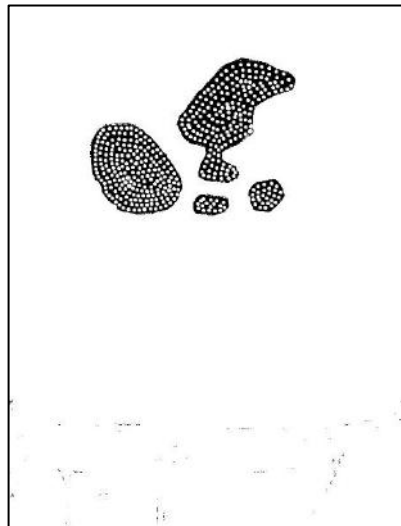
**Figure 5:** The result from the proposed approach

### 3.7. Experimental run

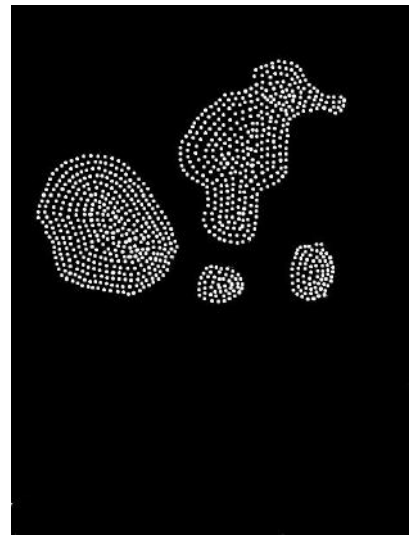
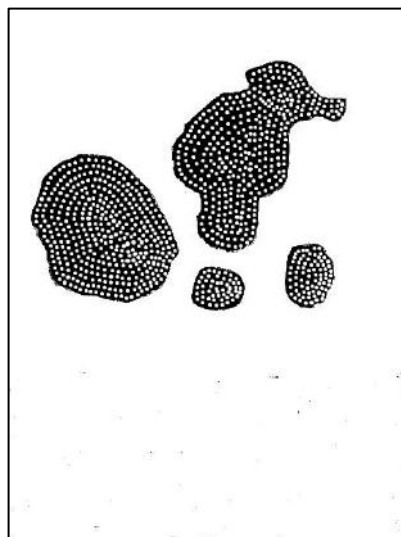
The proposed approach is run on a video for an experimental trial. Three frames are extracted from the video feed and are converted into phosphene images. At first the frames are extracted from the video feed and then the objects in the images are identified with the help of ROI pooling. The identification is done with the help of a pre-trained mask-RCNN model. The objects identified are extracted from the image by the image segmentation. In the study, semantic segmentation is adopted for image segmentation. Then the extracted objects are converted into phosphene image which are then superimposed to recreate the scene. The result obtained from the proposed approach after applying it to the image on the three frames of the video feed is shown in figure 6.



(a) 1<sup>st</sup> frame



(b) 2<sup>nd</sup> frame



(c) 3<sup>rd</sup> frame

**Figure 4:** Frames extracted for a video feed to simulate the optical flow estimation and their corresponding phosphine images

#### 4. Conclusion

The main aim of the study is to develop an effective as well as efficient approach to convert the images into phosphene image that could be interpreted by the bionic eye. The proposed approach is four phase method that reads the image frame by frame from the video feed. Then a pretrained mask RCNN model is employed to identify the objects in the images by ROI pooling. Object identification is helpful to determine the region of interest. Then the objects identified in the image are segregated from the background by semantic segmentation. The object images are then converted into a phosphene images which superimposed to recreate the scene. This process is repeated for each and every frame of the video feed. The proposed approach is highly efficient and effective and can be applied for retinal prosthesis for a bionic eye.

#### 4.1. Future Scope

The approach proposed in the study is at its developing stage and therefore object identification is conducted for very limited amount of objects. Hence in the future the mask RCNN model is to be trained to identify a wide variety of objects. However the proposed approach with its limited resources is highly efficient and effective in nature.

#### Acknowledgement:

The authors would like to express their heartfelt gratitude towards the experts of Delhi public School Kuwait and On My Own Technology Pvt. Ltd. The authors would like to inform that the project is not funded by any organization in any form.

#### References:

1. Zrenner, E. (2002). Will retinal implants restore vision?. *Science*, 295(5557), 1022-1025.
2. Sharmili, N., Swapna, N., & Ramakrishna, G. (2017, April). Comparative analysis of image processing algorithms for visual prosthesis. In *2017 International Conference on Communication and Signal Processing (ICCSP)* (pp. 1120-1124). IEEE.
3. Beyeler, M., Boynton, G. M., Fine, I., & Rokem, A. (2017). pulse2percept: A Python-based simulation framework for bionic vision. *BioRxiv*, 148015.
4. Ayton, L. N., Barnes, N., Dagnelie, G., Fujikado, T., Goetz, G., Hornig, R., ... & Petoe, M. A. (2020). An update on retinal prostheses. *Clinical Neurophysiology*, 131(6), 1383-1398.
5. Han, N., Srivastava, S., Xu, A., Klein, D., & Beyeler, M. (2021, February). Deep learning-based scene simplification for bionic vision. In *Augmented Humans Conference 2021* (pp. 45-54).
6. Yuan, J. C. C., Kaste, L. M., Lee, D. J., Harlow, R. F., Knoernschild, K. L., Campbell, S. D., & Sukotjo, C. (2011). Dental student perceptions of predoctoral implant education and plans for providing implant treatment. *Journal of dental education*, 75(6), 750-760.
7. Walny, J., Carpendale, S., Riche, N. H., Venolia, G., & Fawcett, P. (2011). Visual thinking in action: Visualizations as used on whiteboards. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2508-2517.
8. Pérez L, Rodríguez Í, Rodríguez N, Usamentiaga R, García D, et. al. Robot guidance using machine vision techniques in industrial environments: a comparative review. *Sensors*. 2016;16:335.
9. Wang, J., Zhu, H., Liu, J., Li, H., Han, Y., Zhou, R., & Zhang, Y. (2021). The application of computer vision to visual prosthesis. *Artificial Organs*, 45(10), 1141-1154.
10. Ilea, D. E., & Whelan, P. F. (2011). Image segmentation based on the integration of colour-texture descriptors—A review. *Pattern Recognition*, 44(10-11), 2479-2501.
11. Sanin, A., Sanderson, C., & Lovell, B. C. (2012). Shadow detection: A survey and comparative evaluation of recent methods. *Pattern recognition*, 45(4), 1684-1695.
12. Chakraborty, A., Staib, L. H., & Duncan, J. S. (1996). Deformable boundary finding in medical images by integrating gradient and region information. *IEEE Transactions on Medical Imaging*, 15(6), 859-870.

13. Delahoz, Y. S., & Labrador, M. A. (2014). Survey on fall detection and fall prevention using wearable and external sensors. *Sensors*, *14*(10), 19806-19842.
14. Wang, J., Wu, X., Lu, Y., Wu, H., Kan, H., and Xinyu, C. Face recognition in simulated prosthetic vision: Face detection-based image processing strategies. *Journal of neural engineering* *11* (06 2014), 046009.
15. Guo, F., Yang, Y., Xiao, Y., Gao, Y., and Yu, N. Recognition of moving object in high dynamic scene for visual prosthesis. *IEICE TRANSACTIONS on Information and Systems* E102-D (2019), 1321 {1331.
16. Han, N., Srivastava, S., Xu, A., Klein, D., and Beyeler, M. Deep learning-based scene simplification for bionic vision. In *Augmented Humans Conference 2021 (New York, NY, USA, 2021)*, AHs'21, Association for Computing Machinery, pp. 45-54.
17. Sanchez-Garcia, M., Martinez-Cantin, R., and Guerrero, J. J. Semantic and structural image segmentation for prosthetic vision. *PLoS ONE* *15* (2020).
18. Waldrop, M. M. (2019). What are the limits of deep learning?. *Proceedings of the National Academy of Sciences*, *116*(4), 1074-1077.
19. Qin, Y., He, S., Zhao, Y., & Gong, Y. (2016, November). RoI pooling based fast multi-domain convolutional neural networks for visual tracking. In *2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016)* (pp. 198-202). Atlantis Press.