Semantic and Structural Image Segmentation for Prosthetic Vision

Melani Sánchez-García, Rubén Martínez-Cantín, Jose J. Guerrero

Grupo de robótica, percepción y tiempo real (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: mesangar@unizar.es

Abstract

We present a new approach to build a schematic representation of indoor environments for phosphene images. The proposed method combines a variety of convolutional neural networks for extracting and conveying relevant information about the scene such as structural informative edges of the environment and silhouettes of segmented objects. Experiments were conducted with normal sighted subjects with a Simulated Prosthetic Vision system.

Introduction

Retinal degenerative diseases such as retinitis pigmentosa and age-related macular degeneration cause loss of vision due to the gradual degeneration of the sensory cells in the retina. Visual prosthesis are currently the most promising technology to improve vision in patients with such degenerative diseases. These devices elicit visual perception by electrically stimulating retina cells. As a result, implanted patients are able to see patterns of spots of light called phosphenes that the brain interprets as a visual information [1]. However, retinal implants are limited to hundreds of electrical receptors, which produce a very limited visual elicitation. From the actual technologies for retinal implants, one of the most active line of research is based on implants with a micro camera that captures external stimuli and a processor that converts the visual information in microstimulations in the implant. Prosthetic vision allows users to recognize objects with simple shapes, to see people's silhouettes in bright light or detect motion, but high level tasks require more precise visual cues and a deeper interpretation of the information [1].

Recent developments in implants might result in an improved resolution and performance of the visual elicitation, but quality would still be several orders of magnitude lower than a current micro camera. Alternatively, the visual information gathered by the external camera could be processed prior to being transferred to the retinal electrodes. Image processing can be used to extract and highlight relevant information from the external camera. This information can be presented with visual cues that help to understand the perceived scene by the implanted subject.

In this work, we use semantic segmentation to enhance the visual stimuli for accurate indoor scene understanding using visual prosthesis. Concretely, we use two different types of semantic segmentation based on Fully Convolution Networks (FCNs) to highlight the information available in the image and to present the most useful information to the user. We use instance-aware semantic segmentation to group the pixels of relevant objects in the scene. One of the main problems of using object silhouettes for recognition is the lack of sense of scale or perspective [2]. Thus, we also rely on a second semantic segmentation network to extract structural informative edges of the scenes, such as wall and ceiling intersections. We evaluate and compare the proposed semantic and structural image segmentation with baseline methods (Edge and Direct) through a Simulated Prosthetic Vision (SPV) experiment, which is a standard procedure for non-invasive evaluation using normal vision subjects. The experiments included two tasks: object recognition and room identification.

Stimuli

An overview of the proposed algorithm is shown in Figure 1. Initially, we perform instance segmentation of objects (OMS) using the architecture of Mask R-CNN [3]. The first part of the network, called a Region Proposal Network (RPN), proposes candidates about the regions that contain objects on the input image. The second module, called RoIAlign, runs on the regions of interest (ROIs) proposed by the RPN and aligns those regions to the feature maps extracted by the RPN. Then, the model splits in two branches which generates two outputs for each ROI: the class of the object in the ROI and a bounding box refinement of the object area using a regression model. The mask branch is a convolutional network that takes the positive regions selected by the ROI classifier and generates binary masks. Then, it uses up sampling to

scale the predicted masks to the size of the ROI bounding box which gives the final masks, one per object. Secondly, we extract the structure of the scene by detecting the structural informative edges (SIE), that is, those main edges formed by the intersection of the walls, floor and ceiling of the room. To do that, we use the model by Fernandez-Labrador et al. [4] which is also based on a FCN for pixel classification. Finally, we combine the OMS and SIE as an intelligent way of activating the phosphenes (SIE-OMS). In Figure 2 we show some examples of stimuli used in the experiment.

Conclusions

We present a new visual representation of indoor environments for prosthetic vision, which emphasizes the scene structure and object shapes. By combining the output of two FCN for structural informative edges and object masks and silhouettes, we have demonstrated how different scenes and objects can be quickly recognized even under the restricted conditions of prosthetic vision. Our results demonstrate that our method is well suited for indoor scene understanding over traditional image processing methods used in visual prostheses. The key idea of our current results is that, with only a few significant elements of the scene, it is possible to obtain a good perception of the environment, even in complex and occluded scenes. (Supported by Projects DPI2015-65962-R (MINECO/FEDER, UE) and BES-2016-078426 (MINECO)).

REFERENCES

[1]. CHEN SC, SUANING GJ, MORLEY JW, LOVELL NH. Simulating prosthetic vision: I. Visual models of phosphenes. *Vision Research.* 2009;49(12):1493–1506.

[2]. SANCHEZ-GARCIA M, MARTINEZ-CANTIN R, GUERRERO JJ. Indoor Scenes Understanding for Visual Prosthesis with Fully Convolutional Networks. *In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications; 2019.*

[3]. HE K, GKIOXARI G, DOLLAR P, GIRSHICK R. Mask r-cnn. *In: IEEE International Conference on Computer Vision; 2017. p. 2980–2988.*

[4]. FERNANDEZ-LABRADOR C, FACIL JM, PEREZ-YUS A, DEMONCEAUX C, GUERRERO JJ. PanoRoom: From the Sphere to the 3D Layout. *arXiv preprint arXiv:*180809879. 2018.

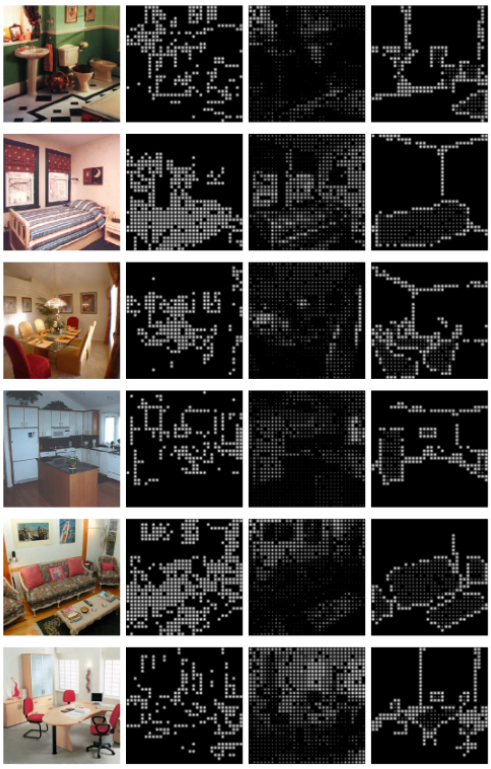

Figure 2: Examples of stimuli used in the experiment. Input image, Edge image, Direct image and SIE-OMS image, respectively.

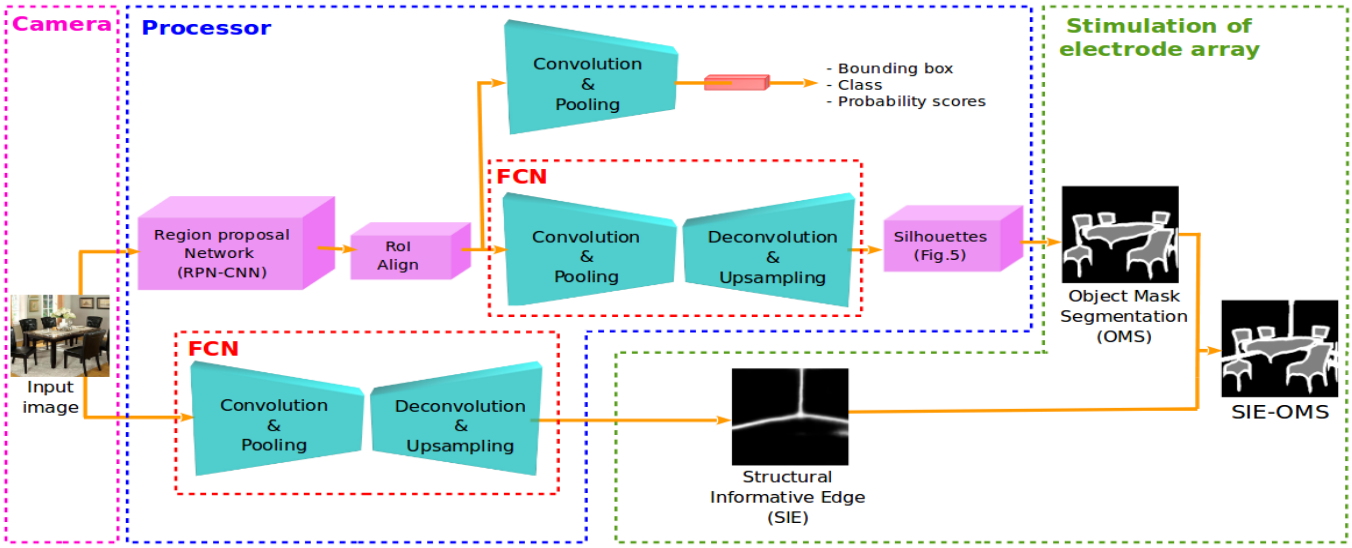

Figure 1: The stimulation of the electrode array is based on two information pathways to extract the regions of pixels that represents important objects (OMS) and structural edges (SIE). The regions are computed using two different types of FCN.