Smart Representation of Indoor Scenes under Simulated Prosthetic Vision

Melani Sanchez-Garcia[1], Ruben Martinez-Cantin[1,2,3], Jose J. Guerrero[1]

[1] I3A, Universidad de Zaragoza, Spain
[2] Centro Universitario de la Defensa, Zaragoza, Spain
[3] SigOpt Inc., San Francisco, CA
(mesangar, josechu.guerrero,rmcantin)@unizar.es

Abstract. Simulated Prosthetic Vision (SPV) is a promising new technology to better understand the perceptual and psychophysical aspects of prosthetic vision. Under prosthetic vision, blind people have a reduced visual perception of the scene. Most of the researches in SPV are based in low level processing of the scene to enhance the perceptions preserving the relevant content of the image. Here, we propose a new approach to build a smart representation of indoor environments for phosphenic images. Our smart representation relies on two parallel CNNs for the extraction of structural informative edges of the room and the relevant object silhouettes based on mask segmentation. We have performed a study with twelve normally sighted subjects to evaluate how our methods were able to the room recognition by presenting phosphenic images and videos. We show how our method is able to increase the recognition ability of the user from $\sim 75\%$ using alternative methods to 90% using our approach.

Keywords: Computer vision, Deep learning, Visual prosthesis , Simulated prosthetic vision

1 Introduction

Implantation of prostheses have become a way to stimulate the surviving neurons in the retina to restore blind vision. In this way, blindness people can perceive simple patterns in the form of spots of light called "phosphenes" (see Fig.1).

However, there are currently some restrictions because of the limit number of implantable electrodes, leading to a low resolution visual perception. Accordingly, it is necessary to optimize the content of the image to aid prosthesis wearers to perform better in visual tasks. This requires using high-level processing to extract more information from the scene and present it to the subject with the phosphenes limitations. Simulated Prosthetic Vision (SPV) become a way to estimate the aspects of prosthetic vision to improve the perceptual quality of the prosthetic wearer, avoiding complex trials in patients. Most of the research used many image processing strategies based on low level processing of the image. For instance, it has been applied for enhancing the scene structure [1].

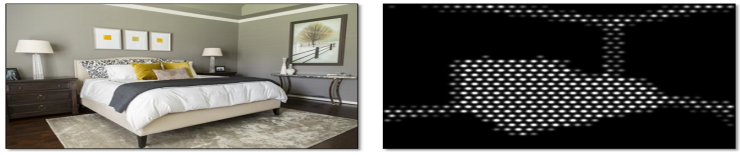

Fig. 1. From an input to output. Left: Original image. Right: Simulated prosthetic vision with 1024 phosphenes

As a new approach of high level image processing, deep learning has revolutionized the field of computer vision [2]. Various applications of diverse range of deep learning such as convolutional neural networks (CNNs) have been used for different tasks such as object detection [3], layout estimation [4] and semantic segmentation [5]. Motivated by the above deep learning approaches, we propose to carry out smart representations of environments in simulated prosthetic vision by utilizing CNNs architectures.

In this work we present a new approach of phosphenic image and video generation based on the combination of relevant object detection and segmentation and the detection of structural edges in indoor scenes. Both the object segmentation and the structural edges are extracted using CNNs, allowing real time processing for video. We evaluate our smart representations for scene recognition both with static images and video.

2 The proposed methodology

In order to address the smart representation of indoor environments, we focus mainly on two parallel neural networks. Concretely, we use a CNNs [6] for the extraction of the structural informative edges of the room and another CNNs [7] based for the segmentation and later the masks generation of relevant objects. Using the output of this two algorithms, we propose two strategies for phosphenic image generation to enhance the limited visual perception under SPV: *Object Mask* (OM), and *Structural Informative Edges combined with Object Masks* (SIE-OM) (see Fig.2). After that, we reduce the resolution of images to 32 x 32.

2.1 Object mask (OM)

OM method is a smart representation composed by a segmentation mask of the most relevant objects in the scene. To carry out this, we use the Mask R-CNN [7]. This is composed by a convolutional neural network (CNN), an extension of Faster R-CNN [8]) used for object detection.

The novelty of Mask-RCNN is a third network, a Fully Convolution Network (FCN), that is applied to each region of interest to perform pixel-wise classification to extract the segmentation masks of each object instance using various blocks of convolution and max pool layers. Thus, the objects silhouettes masks

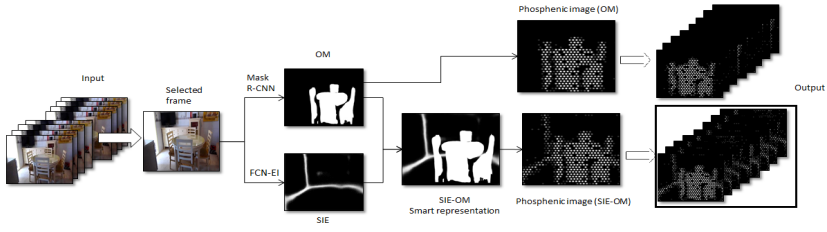

Fig. 2. From an input of frames sequence we use two neural networks algorithms to enhance relevant features in a selected frame to carry out two smart representations, OM and SIE-OM

Table 1. Responses of static images (Ima) and videos (Vid) using OM and SIE-OM methods on object identification and room type recognition tasks

Method	Object present		Object missing		% Correct object identification	% Room type recognized	% Level of Confidence				
	% C	% I	% C	% I			DY	PY	M	PN	DN
OM Ima	11	8	63	18	74 ± 5.06	46 ± 14.81	8	25	29	17	21
OM Vid	14	6	67	13	81 ± 7.13	67 ± 19.32	33	33	13	4	17
SIE-OM Ima	12	6	65	17	77 ± 4.80	54 ± 13.14	8	25	27	21	19
SIE-OM Vid	23	2	67	8	$\mathbf{90} \pm 4.63$	$\mathbf{79} \pm 12.78$	13	67	16	4	0

extracted are used to create the smart representation (OM) of the indoor scene. The rest of the information present in the scene is removed.

2.2 Structural Informative Edges combined with Object Masks (SIE-OM)

The smart representation of SIE-OM is derived from OM method but we added the structural informative edges extracted with [6]. Similarly to the pixel classifier for the object masks described in Section 2.1, this method is also based on a FCN for pixel classification, where given an image, the FCN determines the informative edge map of an image. The result is a binary mask of each of the three types of edges (wall/wall, wall/ceiling, wall/floor) in the layout of the room.

3 Experimental setup and results

We compare the two methods present in Section 2 in two different ways: using *static images* and using *videos*, by measuring the ability of identifying and recognizing several objects and rooms types for a set of subjects using the SPV. For the experiment, subjects were seated in front of a monitor and they visualized randomly a sequence of images and videos processed by the two methods.

The results in Table 1 demonstrate the effectiveness of the proposed methods for smart representation in prosthetic vision extracting relevant information of the scene through two CNNs. The results concluded that the SIE-OM method

supposed an increase of success in the room type recognition compared with the OM method. This suggest that given information of structural edges, scene provide useful information for recognize the room. Furthermore, the videos aid to have a greater understanding of perception and location of objects and edges of the entire scene compared with static images, resulting in a higher percentage of success in the task.

4 Conclusions

We have focused on adding a high-level processing layer to the external images captured by the device in order to increase the information that it is transfered to the subject through the prosthesis. For that purpose, we have used two pixel-wise classifiers for relevant information: structural edges in indoor scenes and relevant object masks.

Using video scenes, subjects were able to obtain a higher percentage of success compared with the static images. Furthermore, subjects were able to recognize the scene displayed with more confidence with the movement of the video. Besides, it has been demonstrated that structural edges provide useful information when it comes to recognizing the room. This meant in a higher performance in the SIE-OM method compared to the OM method. In summary, the subject could understand and recognize different indoor environments under the considerations of prosthetic vision.

References

1. Feng, D., McCarthy, C.: Enhancing scene structure in prosthetic vision using iso-disparity contour perturbance maps. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, IEEE (2013) 5283–5286
2. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4) (1989) 541–551
3. Anisimov, D., Khanova, T.: Towards lightweight convolutional neural networks for object detection. In: Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on, IEEE (2017) 1–8
4. Dasgupta, S., Fang, K., Chen, K., Savarese, S.: Delay: Robust spatial layout estimation for cluttered indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 616–624
5. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognition (2017)
6. Mallya, A., Lazebnik, S.: Learning informative edge maps for indoor scene layout prediction. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 936–944
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2980–2988
8. Girshick, R.: Fast R-CNN. arXiv preprint arXiv:1504.08083 (2015)