Indoor Scenes Understanding for Visual Prosthesis with Fully Convolutional Networks

Melani Sanchez-Garcia[1], Ruben Martinez-Cantin[1,2] and Jose J. Guerrero[1]
[1] *I3A, Universidad de Zaragoza, Spain*
[2] *Centro Universitario de la Defensa, Zaragoza, Spain*

Keywords: Image Understanding, Fully Convolutional Network, Visual Prosthesis, Simulated Prosthetic Vision.

Abstract: One of the biggest problems for blind people is to recognize environments. Prosthetic Vision is a promising new technology to provide visual perception to people with some kind of blindness, transforming an image to a phosphenes pattern to be sent to the implant. However, current prosthetic implants have limited ability to generate images with detail required for understanding an environment. Computer vision play a key role in providing prosthetic vision to alleviate key restrictions of blindness. In this work, we propose a new approach to build a schematic representation of indoor environments for phosphene images. We combine computer vision and deep learning techniques to extract structural features in a scene and recognize different indoor environments designed to prosthetic vision. Our method uses the extraction of structural informative edges which can underpin many computer vision tasks such as recognition and scene understanding, being key for conveying the scene structure. We also apply an object detection algorithm by using an accurate machine learning model capable of localizing and identifying multiple objects in a single image. Further, we represent the extracted information using a phosphenes pattern. The effectiveness of this approach is tested with real data from indoor environments with eleven volunteers.

1 INTRODUCTION

According to the World Health Organization (WHO) the estimated number of people with moderate to severe visual impairment is 285 million (Pascolini and Mariotti, 2012). Implantable prosthetic vision has become a way to stimulate the surviving neurons in the retina providing a very low visual resolution. These devices aim to create visual sensations known as phosphenes through processing the input from a camera by a computer. As a result, blindness people can perceive simple patterns in the form of a set of phosphenes (Guo et al., 2010). Current commercial implants range from 60 to 1500 electrode arrays, which corresponds to pseudo-images of less than 40 by 40 *pixels*. Furthermore, the stimulation of those electrodes is limited in the sense that only few intensity levels can be identified by the user with slow stimulation and recovery times (Luo and Da Cruz, 2014).

Although progress has been made with these prostheses, there are currently some restrictions such as limited visual acuity due to biological challenges, manufacturing constraints and physical restraints (Eiber et al., 2013). The low bandwidth of these prostheses causes the loss of visual information perceived like color, texture, brightness and edges, which is essential for blind people to carry out daily tasks such as navigate and walk safely (Vergnieux et al., 2014) or object recognition and localization (Macé et al., 2015). Thus, it becomes of paramount importance to encode the input information of the environment properly in a way that it can be transmitted efficiently to the subject through the phosphene array.

Following the standard practice in the literature, we use Simulated Prosthetic Vision (SPV) to evaluate the efficacy of visual perceptions in visual prostheses. We also assume that the device includes an external sensor, i.e. a camera, to perceive the scene. Most of the current approaches of SPV use image processing techniques to enhance the perceptions preserving the relevant content of the image (Ayton et al., 2013). This may help to identify regions of interest and highlight the location of objects in the scene. Since only hundreds of pixels are available, both the contrast difference and the edges become very important in terms of the details to be represented in the image.

For image understanding and computer vision, deep learning has recently revolutionized the field

Sanchez-Garcia, M., Martinez-Cantin, R. and Guerrero, J.
Indoor Scenes Understanding for Visual Prosthesis with Fully Convolutional Networks.
DOI: 10.5220/0007257602180225
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 218-225
ISBN: 978-989-758-354-4

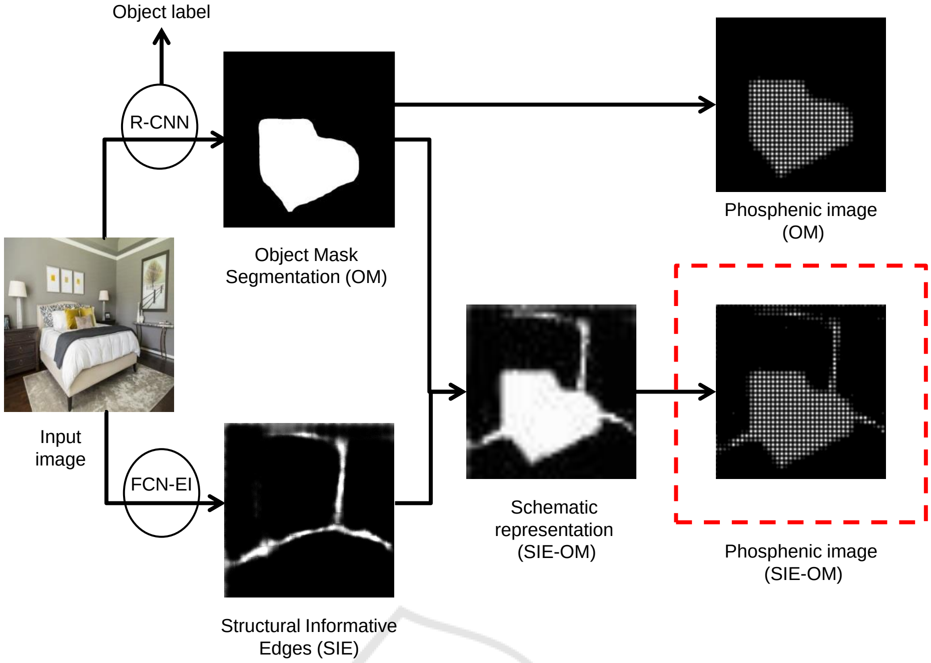

Figure 1: An overview of our approach. We present a new approach of image processing for Simulated Prosthetic Vision. From an input image we used Deep Learning algorithms to enhance relevant features in a scene. Structural Informative Edges (SIE) are generated from (Mallya and Lazebnik, 2015) and Objects Masks Segmentation (OM) are generated from (He et al., 2017). The extracted information is superimposed (SIE-OM) to subsequently create the phosphenic image.

from the seminal work of (LeCun et al., 1989) to the groundbreaking results of (Krizhevsky et al., 2012). Concretely, convolutional neural networks (CNNs) have achieved state of the art results in different tasks such as object detection (Anisimov and Khanova, 2017), layout estimation (Dasgupta et al., 2016) and semantic segmentation (Gu et al., 2017). The use of deep networks could suppose an advance in SPV since they can be used to carry out representations of environments, leading to large improvements in visual perceptions restored by current low-resolution implants (Sanchez-Garcia et al., 2018).

The contribution of this paper is twofold. First, we present a new approach of phosphenic image generation based on the combination of relevant object detection and segmentation and the detection of structural edges in indoor scenes. Our image processing pipeline is shown in Figure 1. Second, we performed an experiment with multiple subjects to validate the performance of our proposal for indoor scene recognition. We also compared the performance of the subjects with a classical method to obtain low bandwidth information such as edge detection (see Figure 2).

2 PHOSPHENE GENERATION USING DEEP LEARNING

2.1 Phosphene Generation Strategies

In this work, we focused on the schematic representation of indoor environments through deep networks. The goal is to improve the ability to understand unfamiliar environments for simulated prosthetic vision. However, the limitation of the low resolution presented by the visual prosthesis means that we have to focus on extracting the most significant information such as structural edges or simple objects.

We opted to extract the structural informative edges of indoor scenes that are informative about the layout of a room. These structural edges are those formed by the intersections of room faces (two walls, wall and ceiling, wall and floor). To accomplish this, we choose a fully convolutional network proposed in (Mallya and Lazebnik, 2015) for pixel-wise labeling that learn to predict 'structural informative edges' for different structures of indoor scenes. We also used a new convolutional network approach proposed in (He et al., 2017) focused on the segmentation of object instances in order to extract more information from

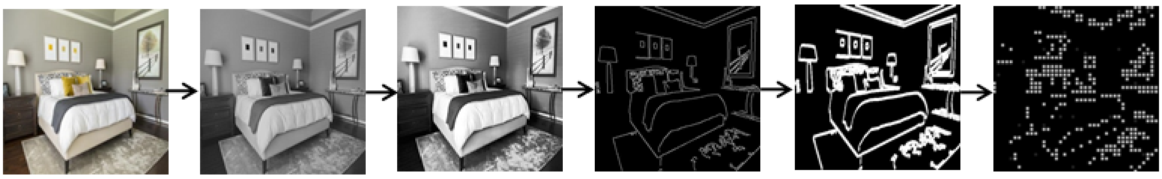

Figure 2: Steps of image processing by using classical Canny method. a) Input image, b) contrast and brightness enhancement c) grayscale histogram equalization, d) canny edge detection, e) image expansion, f) phosphenic image.

the scene through segmentation and later the masks generation of relevant objects. After structural informative edges and object segmentation masks are extracted, we further propose two image enhancement strategies: - Object mask (OM), and a schematic representation formed by - Structural Informative Edges combined with Object mask (SIE-OM) to enhance the limited visual perception under SPV (Figure 1).

2.2 Structural Informative Edges

In the literature exist different approaches for recovering and predicting the layout of an indoor scene (Hedau et al., 2012; Schwing et al., 2012). Our idea was centered on finding edges that are very informative about the structure of a room. Unfortunately, most of these edges are not always visible since the presence of objects in the room occlude them. The neural network proposed in (Mallya and Lazebnik, 2015) is based on a Fully Convolutional Network (FCN) where given an image, it determines the informative edge map of the image. The idea of using this CNN to predict informative edges motivates us to extract directly structural edges of rooms as relevant information to understand the layout of the room in a schematic way. For training this FCN, they use the dataset from (Hedau et al., 2009) which achieve state-of-the-art results on the LSUN dataset (Mallya and Lazebnik, 2015). This network has the particularity that it is trained for two joint tasks. The first task is based on the prediction of the informative edge map generating a binary mask of each of the three types of edges (wall/wall, wall/ceiling, wall/floor) in the layout. The second trained task is based on the prediction of geometric contexts labels. This labels correspond to different room faces plus an object or cluster label. In this FCN, the total loss is considered as the sum of the two cross-entropy classification losses: for the informative edge label prediction and for the geometric context label prediction. Figure 3 (left center) shows examples of structural informative edges detected on different indoor scenes.

2.3 Object Mask Segmentation

We also use the Mask R-CNN method (He et al., 2017) to extract a segmentation mask of objects. Mask R-CNN goes beyond with the novelty of providing an object mask after its segmentation, which is interesting to select and extract objects from the images to later represent it under simulated prosthetic vision. Its approach is summarized in two parts; the first is centered on a convolutional network used for the extraction of features on an entire image, and the second corresponding to the network head for the recognition of bounding box (classification and regression) and mask prediction. This is illustrated in Figure 3 (center).

For this work, we developed two image processing strategies for the input scenes.

2.3.1 Object Mask (OM)

For this strategy, we started making a selection of objects that we considered most relevant in each type of room to subsequently segment and extract a mask for each of them by using the framework of (He et al., 2017). The rest of the information present in the scene were removed. Figure 3 (center) shows examples of the extracted masks from the segmentation of the different objects selected for each type of room. The objects masks extracted from the original images are used to create a schematic representation of the indoor scene. Then, the pixels were reduce to 40x40 and they were represented by simple phosphenes pattern using the approach considered in (Bermudez-Cameo et al., 2017) (see Figure 4).

2.3.2 Structural Informative Edges and Object Mask (SIE-OM)

Here, we proposed a new image processing method based on two neural networks (Mallya and Lazebnik, 2015; He et al., 2017). This method is derived from OM method but we added the structural informative edges extracted with (Mallya and Lazebnik, 2015). In the same way as in the previous method, the pixels were reduce to 40x40 and they were represented by the phosphenes pattern. Figure 3 (right center) shows

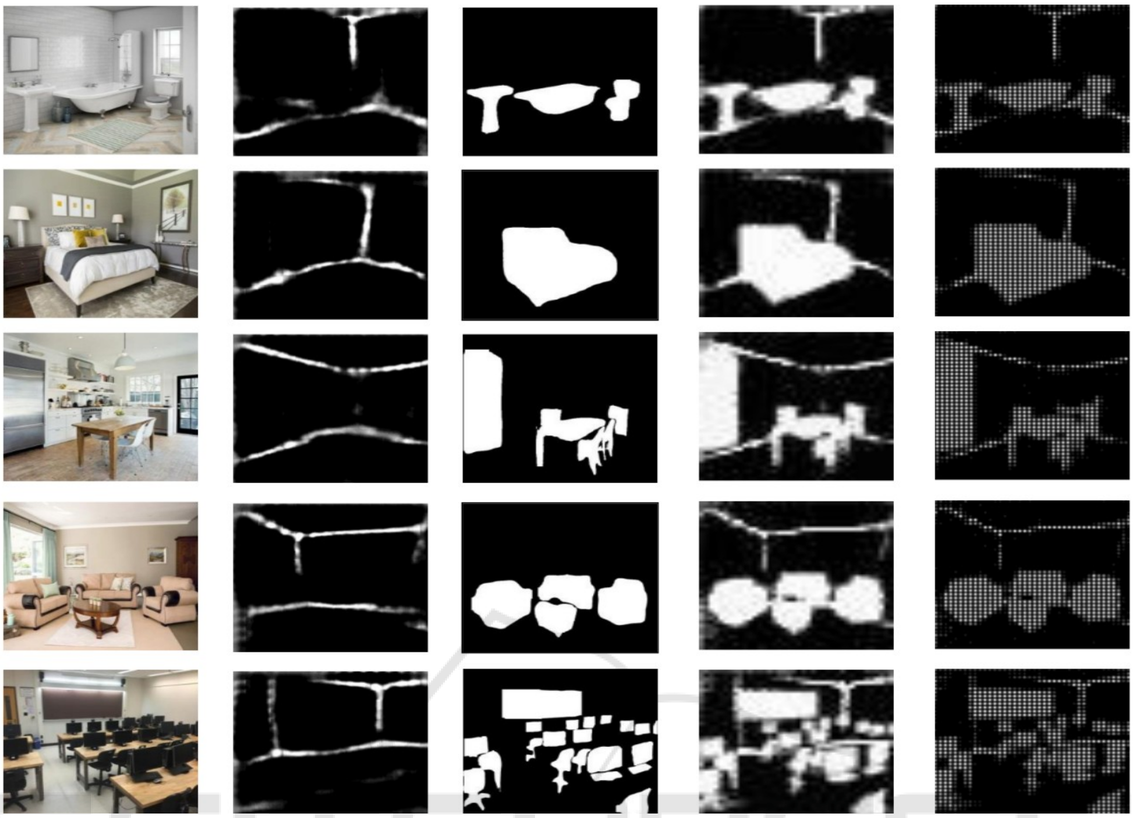

Figure 3: Each row shows an example of indoor scene. The first column shows the input image. The next column shows the structural informative edges of the room. Column three shows the objects masks (OM) and in the fourth column the results of the two convolutional neural networks are represented in the same image (SIE-OM). In the last column, the images are represented by a phosphenes pattern.

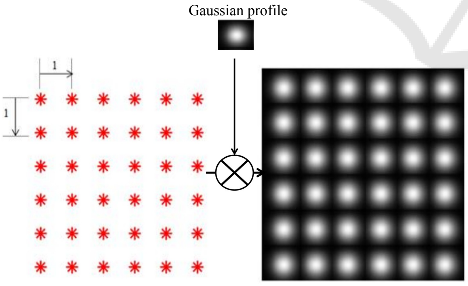

Figure 4: Simulation of phosphene map. Left: Phosphene configuration to create a square grid. Right: Gaussian smoothing applied to the square grid.

examples of schematic representation (SIE-OM) of indoor scenes.

3 EXPERIMENTAL SETUP

3.1 Participants

Eleven subjects (two women and nine men) were volunteered in the experiment. The subjects were between 20 and 36 years old and they were not familiar with the setup and the SPV.

3.2 Ethics Statement

All of the experimental process were conducted according to the ethical recommendations of the Declaration of Helsinki. All subjects were informed about the purpose of the experiment. They could leave the study at any time.

3.3 Phosphene Images

Ten indoor images were collected and were cropped to a size which covered a visual angle of 20° (diagonally) to correspond to epiretinal prostheses (Zhou et al., 2013). Then, the resolution of the images were adjusted to 40x40 *pixels*. We processed five different types of indoor scenes using the image processing strategies from Section 2.1. Similarly to many SPV studies such as in (Chen et al., 2009), phosphenes are approximated as grayscale circular dots with a Gaussian luminance profile (Figure 4). The representation

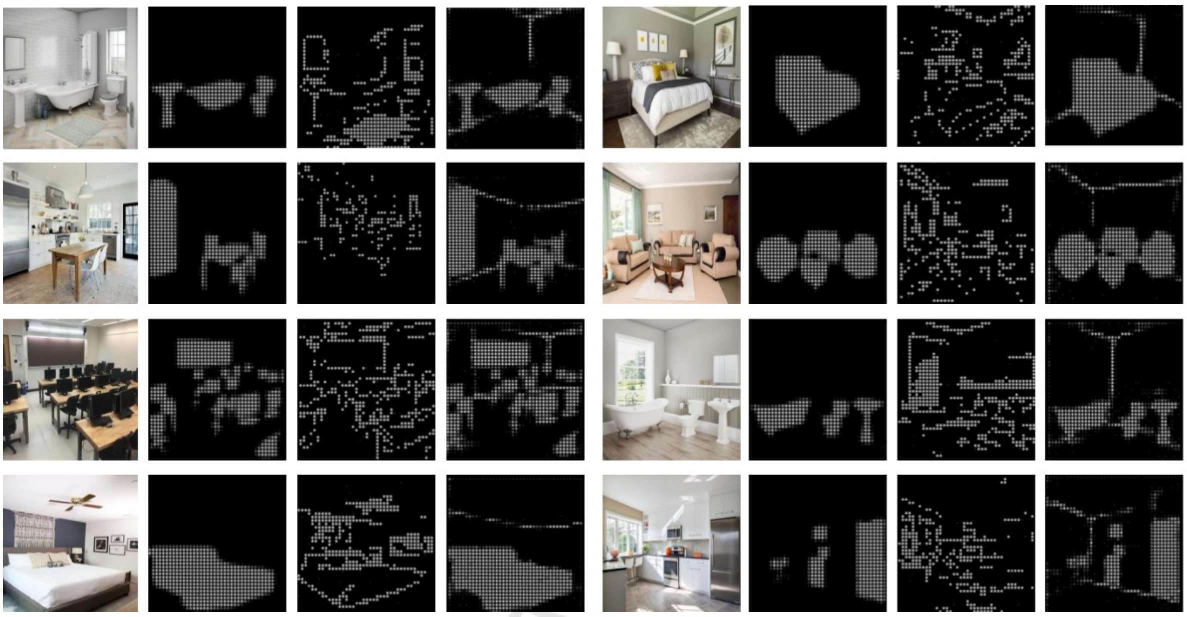

Figure 5: Eight examples of indoor environments represented with 1600 phosphenes. We wanted to show the difference of using a classic method of image processing and a new approaches based on deep learning techniques. Each column show original images, images processed by using OM method, images processed by using Canny method and images processed by our method SIE-OM, respectively.

of the scenes includes six intensity levels of phosphenes, to present the structural informative edges and objects masks in white and gray. Figure 5 gives examples of phosphenic images processed by some of the methods described above (OM and SIE-OM).

3.4 Procedure

Subjects had to perform two different tasks based on object identification and room type recognition with a sequence of images presented on a monitor. A brief introduction of different tasks procedures was given verbally to all the subjects before the experiment. Besides, subjects were informed about the different types of rooms and objects that they were going to see. But they were not informed about the number of objects in each image. The global task of the experiments was to recognize an indoor environment among five types of rooms and the possible objects presents in each scene. Our selection of images included bathrooms, bedrooms, kitchens, living rooms and offices, and the selection of objects was composed by toilets, sinks, baths, refrigerators, ovens/microwaves, tables, chairs, TV's/laptops, beds and couches.

There were two different tasks in the formal tests. The first task was composed by sixteen images of different indoor environments, each one processed in the methods presents in this document. The image presentation sequence was randomized for each subject. Their task were to recognize different objects in the phosphenic images. For this, they had to mark with a cross the boxes in a quiz of the objects identified in each image. The second task was to identify the type of room to which each image corresponds, i.e. they had to determine room types based on what they were seeing on each phosphenic image. Moreover, they had to mark with a cross the certainty with which they were able to recognized it according to the Likert ranking explained later. After each test image, the next image was configured consecutively. The time allotted for each image is limited to 30 seconds. The whole experiment had an average duration of 15 minutes per subject.

4 RESULTS

We compared the two methods present in Section 2.1 with a baseline method using classical image processing tools. This baseline, which we called Canny, is focused on the Canny Edge Detection algorithm that it is typically used for image processing in SPV (Dowling et al., 2004). The performance of this method is based on detecting the edges in an image. Here, the amount of data is significantly reduced, while the important structural properties of the image are preserved. To improve the perception of the information present in the image, we perform six steps for this

Table 1: Comparison of responses of the three different methods (Canny, OM and SIE-OM) on object identification and room type recognition tasks. The assurance of the three methods was evaluated with the Likert ranking.

Method	Object Identified		Object Not Identified		% Correct Object Identification	% Room Recognized	% Level of Confidence				
	%C	%I	%C	%I			DY	PY	M	PN	DN
Canny	7	2	72	17	79	32	0	14	9	23	55
OM	13	5	74	10	87	41	9	32	23	23	14
SIE-OM	13	1	75	10	88	55	27	18	27	5	23

Table 2: Summary of responses of ten images of different indoor environments processed by using SIE-OM method. The performance is evaluated with the percentage of objects and room types correctly and incorrectly identified. For most of the images a high percentage of successes was obtained in the identification of objects. In the case of the room type recognition, the average success result was 58%. The majority of responses were identified with 'DY' and 'PY', which indicates that the subjects understood the visualized scenes.

Image Type	Object Identified		Object Not Identified		% Correct Object Identification	% Room Recognized	% Level of Confidence				
	%C	%I	%C	%I			DY	PY	M	PN	DN
Bathroom_1	23	2	68	7	91	64	64	9	18	0	9
Kitchen_2	13	5	67	15	80	55	0	27	36	18	18
Bedroom_3	10	0	90	0	100	82	55	36	0	0	9
Living room_4	7	0	90	3	97	55	9	45	9	27	9
Office_5	20	1	69	10	89	64	0	55	27	9	9
Bathroom_6	28	0	70	2	98	82	82	18	0	0	0
Kitchen_7	17	2	70	11	87	9	18	45	27	9	0
Bedroom_8	4	6	84	6	87	36	45	36	9	9	0
Living room_9	17	1	79	3	96	82	27	55	9	9	0
Office_10	23	2	70	5	93	55	0	36	36	18	9
Total	16	2	76	6	92	58	30	36	17	10	6

method (see Figure 2). To extract the contours with Canny method we carefully selected the the thresholds for each type of room (from 0.25 to 0.4) to provide optimal performance for each image. In the same way, we selected each line size for image expansion. Note that our approach from Section 2.1 is fully automated requiring no parameter.

The results fall into two main sections: object identification and room type recognition tasks. To obtain the results we collected the percentage of 'Correct' (C) and 'Incorrect' (I) responses in both tasks. We also analyzed the responses with the Likert ranking; the responses marked as 'Definitely yes' (DY), 'Probably yes' (PY), 'Maybe' (M) and 'Probably no' (PN) were assumed as the subject had thought that they had recognized the type of room related to the visualized image. It was assumed that the responses selected as 'Definitely no' (DN) meant that subjects did not understand the type of room witnessed (see Table 1 and Table 2).

4.1 Object Identification

Canny method had the less percentage of objects identified correctly, compared to OM and SIE-OM methods. Besides, Canny images had a 55% of 'DN' responses, which suggests that most of the images were not understood by the subjects. The main reason was perhaps the Canny method extracts too much information that is difficult to understand when the images have low resolution. This confirmed that this method is not very useful at low resolutions. In contrast, there was no significant difference between OM and SIE-OM. This is due to the way to represent the objects in the scene was the same for both methods. Furthermore, this suggests that the representation of the structural informative edges in SIE-OM do not provide much information in the identification of objects.

Looking at the results obtained only for the SIE-OM method, the total percentage of correct object identification task was very high, of 88%. Moreover, the highest percentage of confidence with which they had performed the task is distributed among the response of 'DY' and 'PY', which means that subjects could identify the objects easily.

4.2 Room Type Recognition

SIE-OM method had a higher percentage of correct responses for the room type recognition (55%) than the OM (41%) and the Canny method (32%). We also observed that SIE-OM method had a slightly better performance than the OM method. This suggest that given information of structural edges, scenes provide

Table 3: Confusion matrix results of room type recognition by using SIE-OM method. The results show that some subjects confused room types due to the similarity of the objects present in them. Even so, high precision was obtained for our proposed method.

Actual\Predicted	Bathroom	Kitchen	Bedroom	Living room	Office	Total	Recall
Bathroom	0.91	0.00	0.00	0.09	0.00	1.00	90.91
Kitchen	0.05	0.45	0.10	0.30	0.10	1.00	45.00
Bedroom	0.00	0.00	0.67	0.33	0.00	1.00	66.67
Living room	0.00	0.00	0.00	0.95	0.05	1.00	95.24
Office	0.00	0.00	0.00	0.20	0.80	1.00	80.00
Total	0.96	0.45	0.77	1.88	0.95	5.00	
Precision	94.79	100.00	86.96	50.75	84.42		

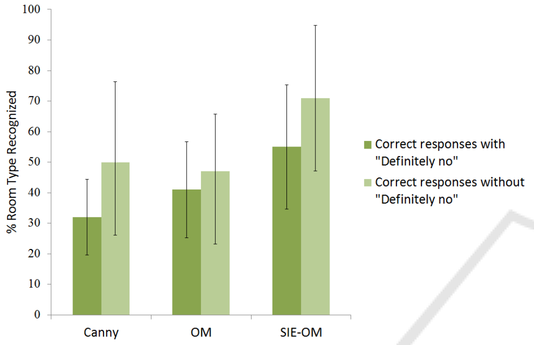

Figure 6: Correct results of room type recognition task obtained by the three methods, with responses of 'DN' and without responses of 'DN'. This suppose an increase in the results of our approach (SIE-OM).

more information when it comes to recognizing the type of room. Figure 6 shows the results on Table 1 compared to the results obtained in the same way but counting as null the responses marked in the Likert ranking as 'DN'. This supposed a significant increase in the successes of the room recognition for SIE-OM method.

Looking at the results of our method, the average percentage obtained of correct responses was 58%. The responses of 'DY' comprised 30% of responses to all images. Instead, responses selected as 'DN' comprised only 6%. This means that subjects could recognize the type of room with a very high certainty with this method (Table 2). In Table 3 is shown the confusion matrix to evaluate the accuracy of our model in the room type recognition. According to the results obtained for the room classification, many subjects confused kitchen with living room. This was due to the presence of tables and chairs in both images. For the same reason, they also confused office with living room. On the other hand, bedroom was confused with living room because the similarity of shape between some beds and couches. Even so, the recall and precision of our model were very high, reaching in some cases 100%.

5 CONCLUSION

We introduced a new approach for simulated visual prosthesis based on computer vision and deep learning algorithms to understand and recognize five types of indoor environments representing the information by simple phosphenes pattern. To address this issue, we used two deep learning techniques proposed by (Mallya and Lazebnik, 2015) and (He et al., 2017) to extract different relevant information from the scenes. To verify this method, we compared it to a classical method of image processing and we studied the effectiveness of the simulated prosthetic vision.

Our overall research has demonstrated that deep learning algorithms can make better use of the limited resolution by highlighting salient features such as structural edges and object segmentation for simulated prosthetic vision. The results show that the new approach proposed has a good performance in the recognition of indoor environments even taking it to a low resolution. As future work we consider enlarge the number of images to evaluate our method with subjects.

ACKNOWLEDGEMENTS

This work was supported by project DPI2015-65962-R (MINECO/FEDER, UE) and BES-2016-078426 (MINECO).

REFERENCES

Anisimov, D. and Khanova, T. (2017). Towards lightweight convolutional neural networks for object detection. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–8. IEEE.

Ayton, L. N., Luu, C. D., Bentley, S. A., Allen, P. J., and Guymer, R. H. (2013). Image processing for visual prostheses: a clinical perspective. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 1540–1544. IEEE.

Bermudez-Cameo, J., Badias-Herbera, A., Guerrero-Viu, M., Lopez-Nicolas, G., and Guerrero, J. J. (2017). Rgb-d computer vision techniques for simulated prosthetic vision. In *LNCS 10255, Pattern Recognition and Image Analysis*, pages 427–436. Springer.

Chen, S. C., Suaning, G. J., Morley, J. W., and Lovell, N. H. (2009). Simulating prosthetic vision: I. visual models of phosphenes. *Vision research*, 49(12):1493–1506.

Dasgupta, S., Fang, K., Chen, K., and Savarese, S. (2016). Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–624.

Dowling, J. A., Maeder, A., and Boles, W. (2004). Mobility enhancement and assessment for a visual prosthesis. In *Medical Imaging 2004: Physiology, Function, and Structure from Medical Images*, volume 5369, pages 780–792. International Society for Optics and Photonics.

Eiber, C. D., Lovell, N. H., and Suaning, G. J. (2013). Attaining higher resolution visual prosthetics: a review of the factors and limitations. *Journal of neural engineering*, 10(1):011002.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2017). Recent advances in convolutional neural networks. *Pattern Recognition*.

Guo, H., Qin, R., Qiu, Y., Zhu, Y., and Tong, S. (2010). Configuration-based processing of phosphene pattern recognition for simulated prosthetic vision. *Artificial organs*, 34(4):324–330.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE.

Hedau, V., Hoiem, D., and Forsyth, D. (2009). Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE.

Hedau, V., Hoiem, D., and Forsyth, D. (2012). Recovering free space of indoor scenes from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2807–2814. IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Luo, Y. H.-L. and Da Cruz, L. (2014). A review and update on the current status of retinal prostheses (bionic eye). *British medical bulletin*, 109(1):31–44.

Macé, M. J.-M., Guivarch, V., Denis, G., and Jouffrais, C. (2015). Simulated prosthetic vision: The benefits of computer-based object recognition and localization. *Artificial organs*, 39(7).

Mallya, A. and Lazebnik, S. (2015). Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 936–944.

Pascolini, D. and Mariotti, S. P. (2012). Global estimates of visual impairment: 2010. *British Journal of Ophthalmology*, 96(5):614–618.

Sanchez-Garcia, M., Martinez-Cantin, R., and Guerrero, J. J. (2018). Structural and object detection for phosphene images. *arXiv preprint arXiv:1809.09607*.

Schwing, A. G., Hazan, T., Pollefeys, M., and Urtasun, R. (2012). Efficient structured prediction for 3d indoor scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2815–2822. IEEE.

Vergnieux, V., Macé, M. J.-M., and Jouffrais, C. (2014). Wayfinding with simulated prosthetic vision: Performance comparison with regular and structure-enhanced renderings. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2585–2588. IEEE.

Zhou, D. D., Dorn, J. D., and Greenberg, R. J. (2013). The argus® ii retinal prosthesis system: An overview. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE.