Smart Representation of Indoor Scenes under Simulated Prosthetic Vision

Instituto Universitario de Investigación en Ingeniería de Aragón
Universidad Zaragoza

Centro Universitario de la Defensa Zaragoza

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

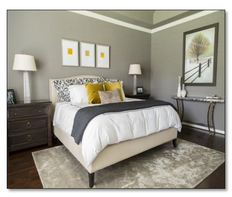
mesangar@unizar.es, rmcantin@unizar.es, josechu.guerrero@unizar.es

[1] Instituto de Investigación en Ingeniería de Aragón (I3A) - Universidad de Zaragoza, [2] Centro Universitario de la Defensa (CUD) - Zaragoza

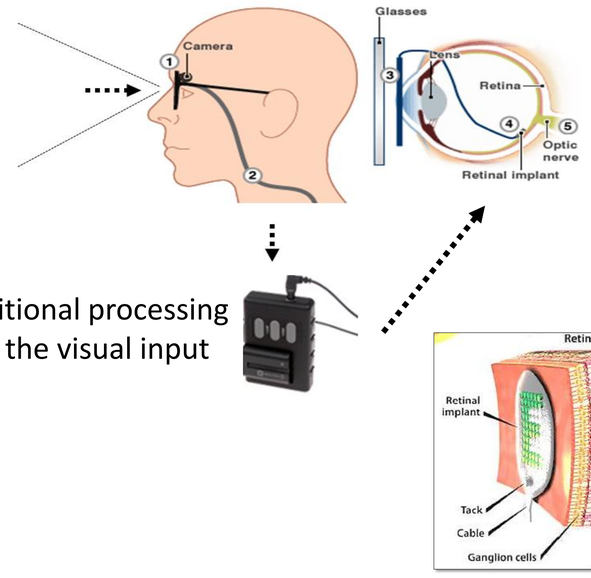
1. Prosthetic vision

Retinal degeneration is caused by the loss of photoreceptors leading to profound blindness

While retinal degeneration destroys the photoreceptors, the neural circuits that convey information from the eye to the brain are sufficiently preserved to make it possible to restore sight using prosthetic devices



Input

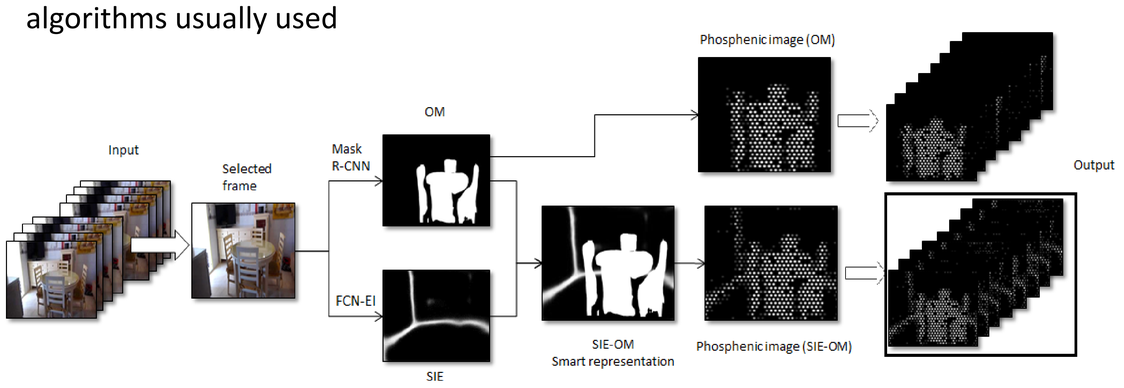


Output: Phosphenic image

Retinal implant

2. Phosphene generation using deep learning

We propose a new strategy to process the visual input information based on high-level image processing algorithms, instead of the low-level image processing algorithms usually used

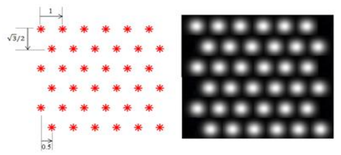


Our smart representation relies on two pixel-wise classifiers of relevant information in indoor scenes:

- **object silhouettes (OM) :** we use a fully convolutional network (FCN) that is applied to each region of interest performing pixel-wise classification to extract the segmentation silhouettes of each object instance

- **structural edges (SIE) :** to represent the structural edges of the room we also use a FCN which is trained for two joint tasks: prediction of the informative edges and geometric context labels

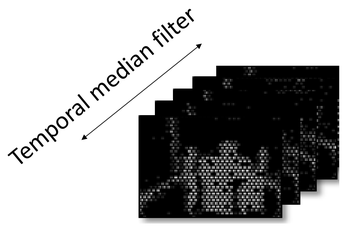
We use a hexagonal phosphene map representing phosphenes as grayscale circular dots with a Gaussian luminance profile



$$G(x,y) \propto \exp\left\{\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma^2}\right\}$$

$$I(x,y) = A(\mu_x,\mu_y)\cdot G(x,y)$$

We compare images and video processed with both methods above. Video introduce motion and a wider field of view which is closer to how a wearer with a retinal implant would see in real life



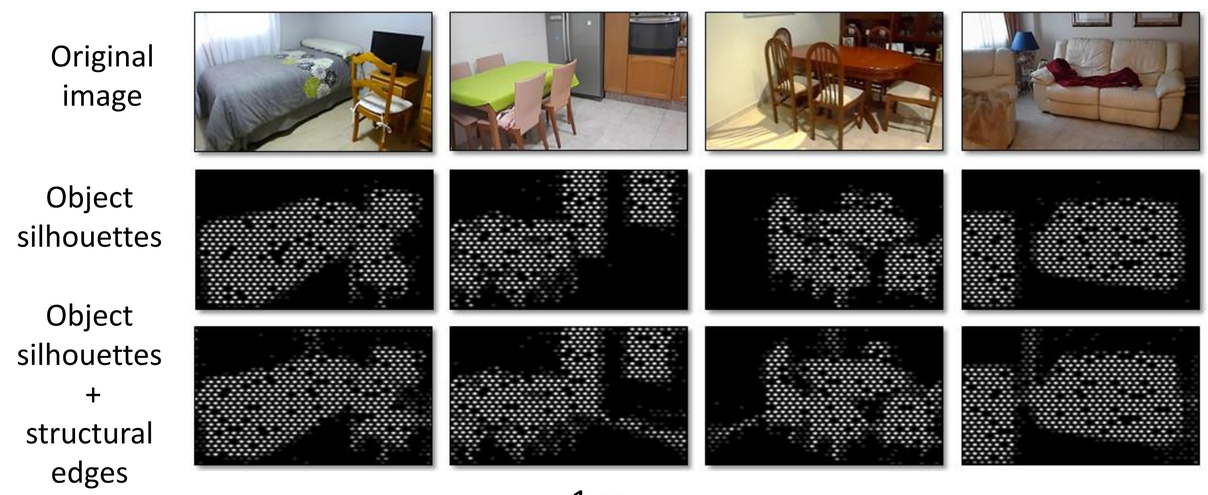
For video, we apply a temporal median filter across five frames which determine each pixels probable value, reviewing which pixel is closest matching to its temporal neighbor

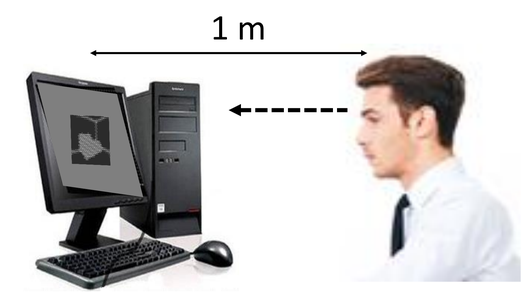
3. Experimental setup

The experiment was carried out with twelve people who have normal sight, evaluating two tasks:

- object identification
- recognition of different indoor rooms

For this the users witnessed 32x32 phosphenic images and video simulations for both methods in front of a computer, while completing a survey

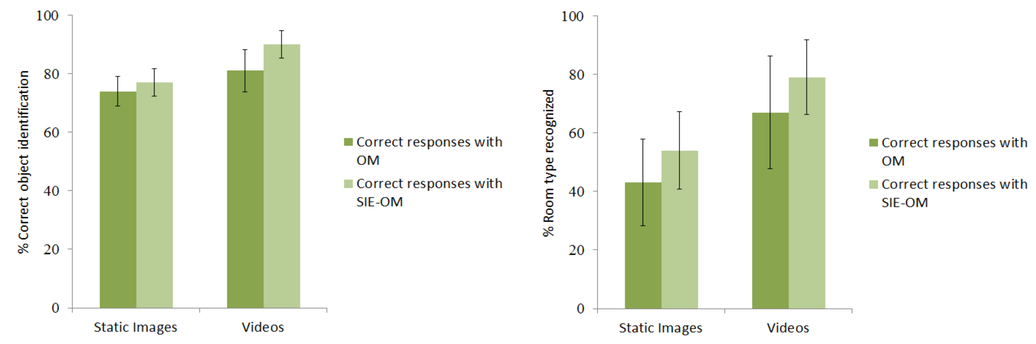




4. Results

We collect the percentage of correct (C) and incorrect (I) responses for the two methods above through images and video. We also analyze the responses with the Likert ranking selected based on the assurance with they had answered the tasks: from total assurance "Definitely yes " (DY) to nothing "Definitely no" (DN)

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



For overall results we include 95% confidence intervals

5. Conclusion

- This is the first method to extract scene information using CNNs for phosphene image generation
- It has been demonstrated that deep learning algorithms can make better use of the limited resolution by highlighting salient features for simulated prosthetic vision
- Results suggested that video increase the performance in the comprehension of the scene compared with images
- Adding structural edges provide useful depth information, which allows a better understanding of the scene

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