Simulated Prosthetic Vision: Improving text accessibility with retinal prostheses

Grégoire Denis, Christophe Jouffrais, Corinne Mailhes and Marc J-M Macé

Abstract — **Image processing can improve significantly the every-day life of blind people wearing current and upcoming retinal prostheses relying on an external camera. We propose to use a real-time text localization algorithm to improve text accessibility. An augmented text-specific rendering based on automatic text localization has been developed. It has been evaluated in comparison to the classical rendering through a Simulated Prosthetic Vision (SPV) experiment with 16 subjects. Subjects were able to detect text in natural scenes much faster and further with the augmented rendering compared to the control rendering. Our results show that current and next generation of low resolution retinal prostheses may benefit from real-time text detection algorithms.**

I. Introduction

Several research groups are designing neuroprotheses which aim at partly restoring vision for blind people. Among these systems, retinal implants are the most advanced (see [1] for a review). To date, two radically different approaches have been proposed: (1) the video images from an external camera are transformed into electrical pulses delivered wirelessly to an internal electrode array, (2) light entering the eye is converted into a pattern of electrical stimulations via micro-photodiodes integrated within the implant. In both devices, electrical stimulations elicit white/yellow spots of light called phosphenes [2]. The resulting crude vision is a set of phosphenes displayed in a restricted portion of the visual field (field of view of 15° to 20°) with little dynamic range (4 to 10 luminance levels per phosphene).

Two groups have been conducting advanced human clinical trials and showed possible benefits for the patients. Argus II (Second Sight Medical Products, Sylmar, CA, USA) is a camera-based retinal prosthesis and consists of a 60-electrode array [3]. It has received FDA and CE approval for commercialization. Alpha-IMS implant (Retina Implant AG, Reutlingen, Germany) is also CE marked and is composed of 1500 micro-photodiodes [4]. Both groups have conducted psychophysical experiments with implanted patients to measure visual task improvements. To date, best observed visual acuity (the sharpness of vision) is 20/1262 and 20/546 with Argus II and Alpha-IMS respectively [1] (legal blindness is below 20/200). Subjects report improvement in tasks such as motion detection, orientation/mobility, and object localization/identification [3]–[5].

All authors are with IRIT Research Laboratory, CNRS & University of Toulouse, Toulouse, France. (Corresponding author: Grégoire Denis, phone: +33 561-556-305; e-mail: elipse_SPV@irit.fr).

Some implanted users are also able to read large white letters (2~40°) and even short words printed on black background [6], [7]. Reading is the most studied psychophysical task under Simulated Prosthetic Vision (SPV) [8]–[14]. These simulators provide a way to experiment prosthetic vision and evaluate new phosphenized renderings with sighted subjects. SPV has shown that hundreds of distinct phosphenes are needed to read at a convenient speed. With the resolution increase brought by next generation implants, reading will probably become more practical but still with large and highly-contrasted letters.

Camera-based implants offer an interesting opportunity to pre-process input images and control phosphenized rendering. It might, for instance, be useful to let the user control image magnification to virtually get closer to a specific target. This scenario makes sense when an implanted person is willing to read a block of text that he cannot get close enough to [14]. However, in this situation, the user has to locate the block of text before zooming or reaching it.

Considering the visual acuity provided by current and next generation implants, most of the implanted users will not be able to localize letters smaller than 2°. With this level of perception, it will be impossible to localize 8cm high letters on public signs (usual letter size for direction indications) when standing further than one or two meters away.

Some recent works focus on task-specific renderings for camera-based implants [15]. In the present study, we propose a rendering based on a text localization algorithm. Here, the input image is processed in order to detect and locate blocks of text in the visual field, and then point out their location to the user. The user may then decide to orient his camera toward a specific block of text, and move closer or zoom in. We designed a SPV system to experiment and evaluate this new rendering. Among all up-to-date text localization algorithms, MSER-based one [16] (Maximally Stable Extremal Regions) has been chosen for real-time implementation capability. Text locations were highlighted by increasing the luminance of phosphenes corresponding to blocks of text.

II. Material

A. Experimental setup

The Virtual Reality headset was a Vuzix 1200AR Head Mounted Display (HMD) (Vuzix Cor., Rochester, CA, USA). Two videos cameras were mounted on the front of the headset to capture the visual scene (640x480 pixels, field of view: 50°, frame rate: 30Hz). The HMD was connected to a

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

computer (dual-core i7) to process the images, perform text localization and compute the augmented phosphene rendering (see Fig. 1B). The phosphene rendering was displayed on the two LCD screens (1024x768 pixels, visual angle: 28×21°, refresh rate: 60Hz) inside the headset. A second computer controlled the experiment progress.

B. Text localization

We chose MSER algorithm to localize text in the images [16]. MSER-type algorithms demonstrated good performances for real-time text detection in natural scenes [17]. Our custom-made software relied on the MSER implementation included in OpenCV library [18]. MSER provided candidate text regions from 640x480 8-bit greyscale input images. All candidates were then filtered (too small or too large regions were discarded, as well as regions with unlikely aspect ratios, e.g. too elongated regions) and grouped together (horizontally and vertically). The final result was a set of regions corresponding to the most likely text locations. The complete processing for one frame took less than 50ms.

C. Simulated Prosthetic Vision

In this SPV experiment, we simulated retinal implants with enough electrodes to restore minimal navigation abilities, i.e. including a few hundreds of electrodes [19], [20]. We simulated two hexagonally settled electrode arrays: a 15x18 and a larger 40x50, subtending 12x17° and 19x25° of visual angle respectively. The larger implant was designed to enable direct reading opportunities. Similarly to many SPV studies [2], phosphenes were approximated as greyscale circular dots with a Gaussian luminance profile. Simulated phosphenes had four luminance levels and a diameter of 0.9° for the smallest array and 0.5° for the largest. Spacing between phosphenes was identical for the two arrays (0.2°). 10% randomly selected dots were switched off to simulate electrode dropout.

We designed two renderings: "standard" and "augmented" (Fig. 1B). In the standard rendering, each video frame was resized (bicubic interpolation) to fit the number of simulated electrodes. The luminance of each simulated phosphene resulted from the luminance of the corresponding pixel in the resized image. The augmented rendering was similar, except that the text localization algorithm was applied on the image to find blocks of text, and highlight them by increasing the brightness of corresponding dots (and decreasing the brightness of the other dots at the same time).

III. Experiment

A. Participants

16 sighted subjects (5 women, 11 men) participated in this study. This experiment was conducted according to the ethical recommendations of the Declaration of Helsinki and was approved by a local ethical committee (CLERIT) at the University of Toulouse. All subjects gave written informed consent to participate.

B. Procedure

Subjects were seated 57 cm away from a TV screen. They had to localize a block of text on natural scenes (street photographs) displayed on the screen (Fig. 1A). Each subject systematically performed four conditions: two array sizes (15x18 and 40x50 simulated phosphenes) in each rendering ("standard" and "augmented"). For each condition, we used a set of 56 pictures (1440x1080 pixels subtending 69x52° of visual angle) where blocks of text were distributed as follows: 8 images without any text, 12 with text in upper left-hand quadrant, 12 in upper right, 12 in lower right, and 12 in lower left. Three font sizes (1°, 2° and 4°) were equally distributed within the 48 pictures containing text.

There were 56 trials per condition, corresponding to the 56 pictures (display order was random). The task was to indicate for each trial the quadrant of the TV screen in which the text was located (if any). Each subject performed a total of 224 trials (56 trials per condition x 4 conditions). The sequence order of the four conditions was counterbalanced across subjects to compensate for potential learning effects. At the beginning of the experiment, subjects were trained during 10 minutes on a dedicated set of pictures. The instruction was to localize as fast and as accurately as possible a text displayed in front of them. A trial started with a short sound after which they were free to move the head (without getting closer) to scan the TV screen and find the text. They gave their responses verbally (0 = no text, 1 = upper left, 2 = upper right, 3 = lower right and 4 = lower left). The whole experiment had an average duration of 1 hour per subject.

IV. Results

Following the experiment, two parameters have been analyzed: the response accuracy (percentage of correct responses), and the response time (time in seconds to provide an answer).

Statistics have been performed with R software 2.15.0[1]. The number of observations was limited, and data distribution was non-Gaussian. Therefore pair-wise group or condition comparisons have been performed with Wilcoxon tests, the significance level for all tests being set at 0.05. Significance values were corrected for multiple pair-wise comparisons using Bonferroni correction.

In the next section, "Std" will refer to "standard" rendering, "Aug" will refer to "augmented" rendering, and SD will refer to standard deviation. Note that if the subject made random decisions between the 5 possible answers for each trial, mean accuracy would be at 20% (chance level).

Table 1 lists subjects' performance (mean accuracy and mean response time) for the 4 conditions. Gathering all the subjects, accuracy was over 90% for the two "Aug" conditions (1518 and 4050). Accuracy decreased significantly to 64.0% (SD=5.8) and 32.9% (SD=7.9) in conditions Std4050 and Std1518 respectively.

[1] http://www.r-project.org

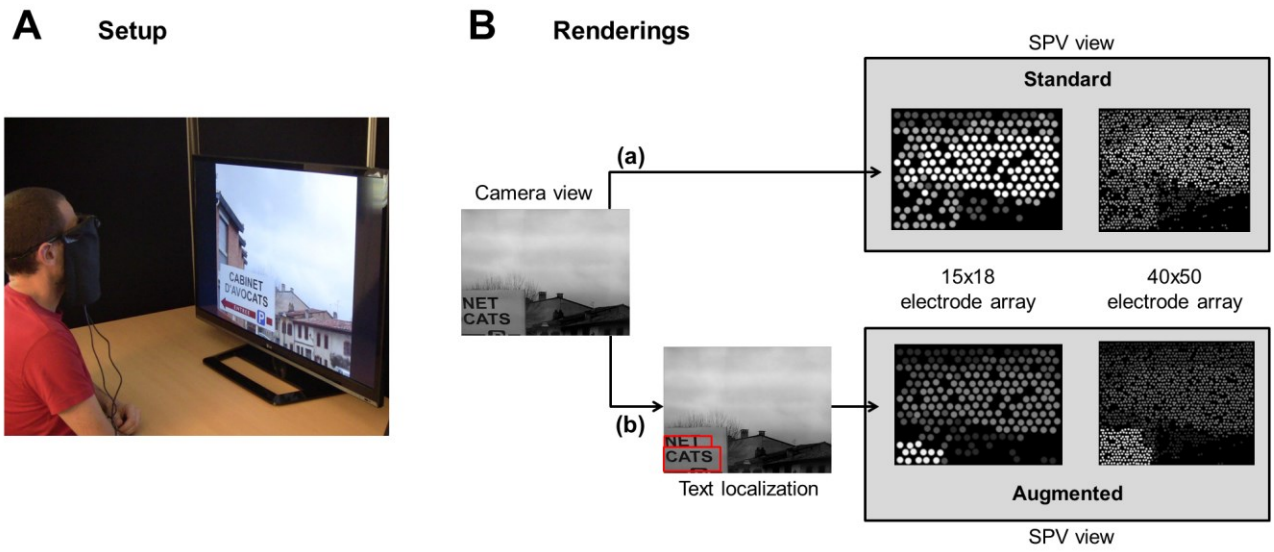

Figure 1. A - Setup overview: subjects were seated 57 cm away from a TV screen and had to localize blocks of text with two different renderings and two different electrode arrays. B - The two renderings: (a) in the standard rendering, image is resized to fit the number of simulated electrodes; (b) in the augmented rendering, blocks of text localized with MSER-based algorithm are highlighted with bright phosphenes.

TABLE I. SUBJECTS PERFORMANCE (ACCURACY AND RESPONSE TIME) AND PAIR-WISE COMPARISONS BETWEEN CONDITIONS (P<0.01 IN LIGHT GREY CELLS)

	Mean	SD	Wilcoxon (Z, p value)			
Accuracy (% correct)			**Std1518**	**Std4050**	**Aug1518**	**Aug4050**
Std1518	32.9	7.9		-3.5	-3.5	-3.5
Std4050	64.0	5.8	-3.5		-3.5	-3.5
Aug1518	90.7	3.9	-3.5	-3.5		-1.3, p=0.2
Aug4050	92.6	3.1	-3.5	-3.5	-1.3, p=0.2	
Response Time (s)			**Std1518**	**Std4050**	**Aug1518**	**Aug4050**
Std1518	12.9	3.8		-3.5	-3.5	-3.5
Std4050	8.9	1.5	-3.5		-3.5	-3.5
Aug1518	6.6	1.5	-3.5	-3.5		-3.1
Aug4050	5.8	1.2	-3.5	-3.5	-3.1	

Pair-wise comparisons between conditions revealed a significant effect on accuracy except between Aug1518 and Aug4050 (Table 1).

For each condition, 8 images had no blocks of text, the correct response being "0". The averaged accuracy (including all conditions) for these "catch" trials was 75.2% (SD=19.9). This confirmed that the subjects were performing the task correctly, even in conditions where text was hard or impossible to perceive.

All subjects together, the average time to successfully localize a block of text was 12.9s in the Std1518 condition, 8.9s in the Std4050 condition, 6.6s in the Aug1518 condition, and 5.8s in the Aug4050 condition. All these response times were significantly different (See Table 1 for details).

In order to roughly assess the effect of distance on text localization, the font size of the letters in the images was controlled (3 groups of 12 images with letters of 1°, 2° and 4°). Fig. 2 reports text detection accuracy according to these different letter sizes. Subjects were at chance level for the 1° letters in the Std1518 condition, and performance slightly improved in this condition with larger font size (32.0% at 4°). With eight times more simulated phosphenes (Std4050), performance was relatively low for 1° letters (36.7%), and increased close to the performance obtained on both "Aug" conditions for the largest letters (89.1% vs. 97.6% & 98.0%). The performance for the two "Aug" conditions was independent from the font size. Indeed, the algorithm was similarly able to detect blocks of texts with letter sizes of 1°, 2° and 4°.

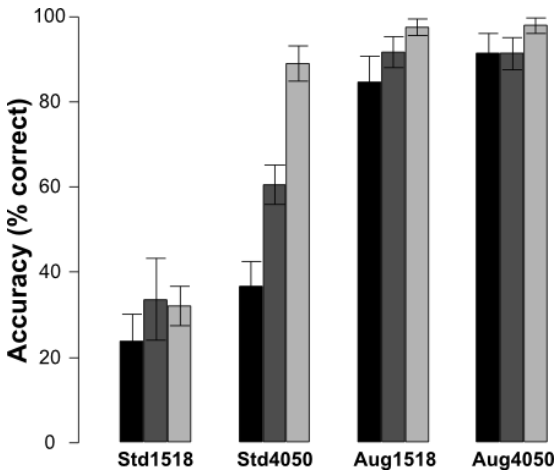

Figure 2. Accuracy (% correct) per condition and letter size. Black: 1°, dark grey: 2°, light grey: 4°.

V. Discussion

In this SPV experiment, we have highlighted that despite intrinsic low resolution, camera-based retinal implants that will be available in the upcoming years could greatly benefit from computer vision techniques. As an example, we have shown that a text detection algorithm could help implanted users detect and locate text in the camera field of view. Indeed, a standard rendering approach with an array of 15x18 electrodes does not allow text detection on street sign at more than 1-2 meters. With the augmented rendering that we proposed, based on the real-time processing of the camera image, it could be possible to localize text at a distance at least 4 to 8 times greater (8 meters at least, but the distance could be increased with higher camera resolution and more processing power).

The MSER-based algorithm used in this study to detect blocks of text was limited both in the size of the letters and the type of font that it could detect. But text detection and recognition algorithms are becoming more and more efficient [17]. Even if mobile processors are not as fast as computer ones, overall power will keep increasing, and a truly generic real-time text detector with a low false alarm rate is doubtlessly feasible within the 5 upcoming years. This achievement could greatly improve text localization, and hence accessibility, to the forthcoming population of blind people implanted with retinal or cortical implants.

As we showed with text detection in the visual field, the inclusion of artificial vision algorithms in camera-based implants opens a large field of everyday life improvement in the usability of prosthetic vision (e.g. faces or objects recognition).It also opens new perspectives on the interaction between the user and the prosthetic device. In the frame of our experiment, a user could intentionally zoom on a detected text, and choose to directly read it, or send it to an OCR (Optical Character Recognition) algorithm and text-to-speech software. More generally, task-specific renderings (navigation, object and face localization and recognition, text localization, etc.) could all be integrated in one device, allowing a prosthesis user to switch between different modes.

VI. Acknowledgments

We would like to thank Jean-Yves Tourneret, professor at the Univ. of Toulouse, IRIT for the discussions concerning text-localization algorithms, as well as Jordi Castillo for his preliminary work on text detection.

References

[1] A. T. Chuang, C. E. Margo, and P. B. Greenberg, "Retinal implants: a systematic review.," *Br. J. Ophthalmol.*, pp. bjophthalmol–2013–303708–, Jan. 2014.

[2] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell, "Simulating prosthetic vision: I. Visual models of phosphenes.," *Vision Res.*, vol. 49, no. 12, pp. 1493–1506, Jun. 2009.

[3] M. S. Humayun, J. D. Dorn, L. da Cruz, G. Dagnelie, J.-A. Sahel, P. E. Stanga, A. V Cideciyan, J. L. Duncan, D. Eliott, E. Filley, A. C. Ho, A. Santos, A. B. Safran, A. Arditi, L. V Del Priore, and R. J. Greenberg, "Interim results from the international trial of Second Sight's visual prosthesis.," *Ophthalmology*, vol. 119, no. 4, pp. 779–88, May 2012.

[4] K. Stingl, K. U. Bartz-Schmidt, D. Besch, A. Braun, A. Bruckmann, F. Gekeler, U. Greppmaier, S. Hipp, G. Hortdorfer, C. Kernstock, A. Koitschev, A. Kusnyerik, H. Sachs, A. Schatz, T. Peters, B. Wilhelm, and E. Zrenner, "Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS," *Proc. R. Soc. B Biol. Sci.*, vol. 280, no. 1757, pp. 20130077–20130077, Feb. 2013.

[5] J. D. Dorn, A. K. Ahuja, A. Caspi, L. da Cruz, G. Dagnelie, J.-A. Sahel, R. J. Greenberg, and M. J. McMahon, "The Detection of Motion by Blind Subjects With the Epiretinal 60-Electrode (Argus II) Retinal Prosthesis.," *Arch. Ophthalmol.*, pp. 1–7, Oct. 2012.

[6] L. da Cruz, B. F. Coley, J. D. Dorn, F. Merlini, E. Filley, P. Christopher, F. K. Chen, V. Wuyyuru, J.-A. Sahel, P. E. Stanga, M. S. Humayun, R. J. Greenberg, and G. Dagnelie, "The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss.," *Br. J. Ophthalmol.*, vol. 97, no. 5, pp. 632–6, May 2013.

[7] E. Zrenner, K. U. Bartz-Schmidt, H. Benav, D. Besch, A. Bruckmann, V.-P. Gabel, F. Gekeler, U. Greppmaier, A. Harscher, S. Kibbel, J. Koch, A. Kusnyerik, T. Peters, K. Stingl, H. Sachs, A. Stett, P. Szurman, B. Wilhelm, and R. G. H. Wilke, "Subretinal electronic chips allow blind patients to read letters and combine them to words.," *Proc. Biol. Sci.*, vol. 278, no. 1711, pp. 1489–97, May 2011.

[8] K. Cha, K. W. Horch, R. A. Normann, and D. K. Boman, "Reading speed with a pixelized vision system.," *J. Opt. Soc. Am. A.*, vol. 9, no. 5, pp. 673–7, May 1992.

[9] J. Sommerhalder, E. Oueghlani, M. Bagnoud, U. Leonards, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision: I. Eccentric reading of isolated words, and perceptual learning.," *Vision Res.*, vol. 43, no. 3, pp. 269–83, Feb. 2003.

[10] J. Sommerhalder, B. Rappaz, R. de Haller, A. Pérez Fornos, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision: II. Eccentric reading of full-page text and the learning of this task.," *Vision Res.*, vol. 44, no. 14, pp. 1693–706, Jan. 2004.

[11] A. Pérez Fornos, J. Sommerhalder, B. Rappaz, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision, III: Do the spatial or temporal characteristics of stimulus pixelization really matter?," *Invest. Ophthalmol. Vis. Sci.*, vol. 46, no. 10, pp. 3906–12, Oct. 2005.

[12] G. Dagnelie, D. G. Barnett, M. S. Humayun, and R. W. Thompson, "Paragraph text reading using a pixelized prosthetic vision simulator: parameter dependence and task learning in free-viewing conditions.," *Invest. Ophthalmol. Vis. Sci.*, vol. 47, no. 3, pp. 1241–50, Mar. 2006.

[13] X. Chai, W. Yu, J. Wang, Y. Zhao, C. Cai, and Q. Ren, "Recognition of pixelized Chinese characters using simulated prosthetic vision.," *Artif. Organs*, vol. 31, no. 3, pp. 175–82, Mar. 2007.

[14] A. Pérez Fornos, J. Sommerhalder, and M. Pelizzone, "Reading with a simulated 60-channel implant.," *Front. Neurosci.*, vol. 5, p. 57, Jan. 2011.

[15] V. Vergnieux, M. J.-M. Macé, and C. Jouffrais, "Wayfinding with Simulated Prosthetic Vision: Performance comparison with regular and structured-enhanced renderings.," in *36th Annual International Conf. of the IEEE EMBS*, 2014.

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, Sep. 2004.

[17] C. Merino-Gracia, K. Lenc, and M. Mirmehdi, "A head-mounted device for recognizing text in natural scenes.," in *CBDAR 2011*, 2012, pp. 29–41.

[18] D. Nistér and H. Stewénius, "Linear time maximally stable extremal regions," in *European Conference on Computer Vision*, 2008, pp. 183–196.

[19] K. Cha, K. W. Horch, and R. A. Normann, "Mobility performance with a pixelized vision system.," *Vision Res.*, vol. 32, no. 7, pp. 1367–72, Jul. 1992.

[20] G. Dagnelie, P. Keane, V. Narla, L. Yang, J. D. Weiland, and M. S. Humayun, "Real and virtual mobility performance in simulated prosthetic vision.," *J. Neural Eng.*, vol. 4, no. 1, pp. S92–101, Mar. 2007.