Human Faces Detection and Localization with Simulated Prosthetic Vision

Grégoire Denis
IRIT – University of Toulouse & CNRS
Université Paul Sabatier
Toulouse - France
gregoire.denis@irit.fr

Christophe Jouffrais
IRIT – University of Toulouse & CNRS
Université Paul Sabatier
Toulouse - France
christophe.jouffrais@irit.fr

Victor Vergnieux
IRIT – University of Toulouse & CNRS
Université Paul Sabatier
Toulouse - France
victor.vergnieux@irit.fr

Marc Macé
IRIT – University of Toulouse & CNRS
Université Paul Sabatier
Toulouse - France
marc.mace@irit.fr

CHI 2013 Extended Abstracts, April 27-May 2, 2013, Paris, France
ACM 978-1-4503-1952-2/13/04.

Abstract
Clinical trials reveal that current visual neuroprosthesis are not yet usable. The main reason is the small number of implanted electrodes, leading to a very poor visual resolution. The resolution is especially not sufficient to detect specific objects (faces, signs, etc.) in the surroundings. We used simulated prosthetic vision (SPV) to show that pre-processing of the camera image could restore these functions, even with low-resolution implants. Specifically, we showed that it is possible to quickly detect and localize human faces located nearby. We suggest that high-level processing of the video stream may be included in current visual neuroprosthesis. This would restore many visuomotor behaviors such as grasping, heading, steering, etc.

Keywords
visual impairment; retinal prosthesis; simulated prosthetic vision; computer vision

ACM Classification Keywords
K.4.2. [Computers and Society]: Social Issues - Assistive Technologies for Persons with Disabilities; H.5.2. [Information Interfaces and Presentation]: User Interfaces; I.4.9. [Image Processing and Computer Vision]: Applications

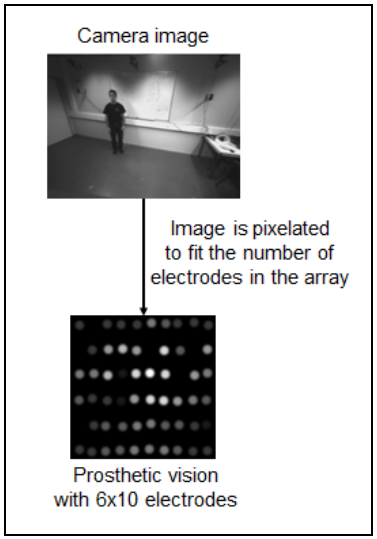

Figure 1 Illustration of prosthetic vision. A blind implanted patient perceives the scene through a camera and stimulating electrodes in the retina or visual cortex. In the illustrated "scoreboard" rendering, the camera image is resized to 6x10 pixels to fit the number of electrodes in the implant. A very low resolution of the visual scene is perceived via the implant.

Introduction

Visual neuroprostheses, consisting of a camera connected to an array of electrodes implanted in the eye or the brain to bypass damaged neural tissues have been developed for the last 30-40 years. A few retinal implants [7,9] are currently undergoing clinical trials prior to commercialization. The resolution of these early implants is still very limited (6*10 electrodes for the most recent clinical trials [7]). A rapid increase in the electrode number is unlikely: arrays with a hundred electrodes were fabricated more than 20 years ago and are still waiting for implantation procedures [2]. Retinal implants evoke visual percepts in the form of white dots called phosphenes. Classically, the resolution of the camera image is drastically reduced to match the number of implanted electrodes. Then, this low resolution version of the visual scene is rendered via the phosphene array (**Figure 1**). This method is sometimes called a "scoreboard" rendering. The low resolution of this rendering is an obstacle to usability, and visual neuroprosthesis are thus effective in a very limited set of situations only [5,10]. Indeed, many visual tasks are still difficult or impossible for implanted blind persons, such as remote objects identification, fluent reading, navigation in unknown environments or detection of surrounding objects or persons.

Simulated prosthetic vision (SPV) has been developed by several research groups. This technique provides an easy way to forestall prosthetic vision problematic and assess new rendering strategies, in the absence of implanted patients. To date, most research work focused on resampling the camera image or digital zooming. However, useful perception always relies on a relatively high number of electrodes/phosphenes [3].

We advocate that functionality and usability of current and upcoming low resolution implants could be greatly improved if a processing was applied on the camera image before rendering it. Indeed, the processing of the image may help to identify regions of interests and/or highlight the location of important objects in the scene. This augmented information could be displayed at the same time as the "scoreboard" rendering to improve visuo-motor behaviors such as orientation, steering, grasping, etc.

To test our hypothesis, we developed a SPV based on the design of an existing implant (Argus II with 6x10 electrodes) and a real-time image processing algorithm. For the sake of experimental validation, we focused on human face detection in the peripersonal space.

In this study, we show that real-time processing of the video with a face detector can be used to add additional information to the "scoreboard" rendering of the scene. This composite display, with the visual scene and additional information on faces could help locate other persons in the surroundings. Based on this "augmented reality"-like approach, low resolution visual prosthesis may help restore many visual behaviors that are not possible with the classical "scoreboard" rendering.

Material and methods

Apparatus

The SPV included a Head-Mounted Display (HMD) and a binocular camera (Bumblebee II - 03S2, Point Grey, USA) with a resolution of 320x240 pixels at a rate of 48 frames per second. The camera viewing angle was 100 degrees. The camera was attached on top of the HMD. Additionally, we used a motion capture system with 12 cameras (OptiTrack, Natural Point, USA) in order to track the subject's hands and shoulders.

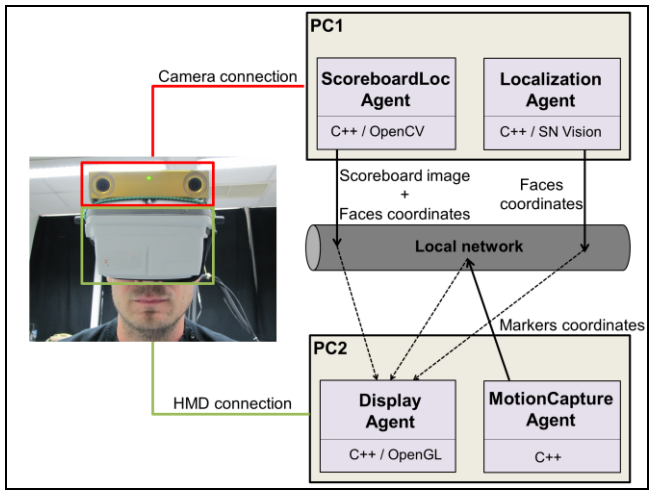

Figure 2 Simulator of Prosthetic Vision: architecture overview. A sighted user is deprived from normal vision. He only perceives white dots mimicking phosphenes in the HMD. The "ScoreboardLoc Agent", handled SCB+LOC condition (image resampling and faces detection). The "Localization Agent" handled LOC condition (faces detection only). The "Display Agent" rendered the phosphenes for the prosthetic vision simulation.

The system architecture (**Figure 2**) was composed of two (quad-core CPU) computers. The first one hosted the real-time image processing (minimal output frequency of 15Hz). The second one managed the implant simulation and the motion capture system.

The simulated phosphenes were presented on an NVisor SX-60 HMD (NVIS Inc., USA) with a resolution of 1280x1024 pixels, subtending 44x34 degrees of visual angle.

Face detection

Face detection was performed with a computer vision bio-inspired algorithm (Spikenet Technology, France). We chose this algorithm because of its robustness to image transformations together with sheer speed [6]. The algorithm looks for the closest matches between the current frame in the video stream and pre-learnt models of target objects. 10 models (50x50 pixels) per face were needed to allow recognition at every position in the room.

Simulated implant design

The simulated implant was a 6x10 electrode array spanning 11° of visual angle. Phosphene appearance and implant design were based on observations acquired during multiple clinical trials of electrical stimulation of the retina [3]:

- phosphenes were roundly shaped with a Gaussian luminance profile,
- phosphenes had 8 levels of luminance,
- phosphenes size was 1°,
- phosphenes were squarely disposed with some noise on exact position,
- 10% randomly selected phosphenes were switched off to simulate electrodes dropout.

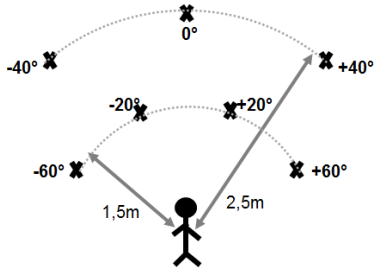

Figure 3 Subjects were asked to determine the number and the position of faces in front of them. In each trial, 0, 1 or 2 persons were placed at a specific location. The seven possible locations were labeled with marks on the ground.

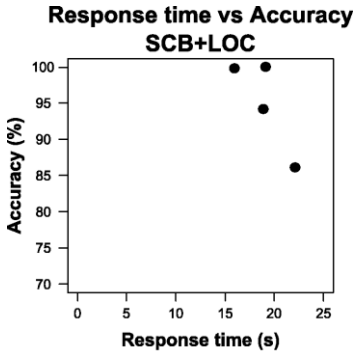

Figure 4 Response time versus Accuracy (SCB+LOC condition only)

In addition, we implemented a specific feature in order to manage rapid adaptation effects observed with long-lasting electrical stimulation (1-2s in the retina). Each phosphene was linked to a countdown timer. A refresh (switch off for 100 ms) occurred whenever the phosphene had kept the same luminance for a specified amount of time (here 2.0 ±0.2 s) [8,10].

Global behavior

We designed two rendering conditions: SCB+LOC (Scoreboard rendering augmented with face localization) and LOC (Localization information only). In the first one, we displayed the scene as in the classical "scoreboard" approach and highlighted the location of the detected faces by changing the appearance of the corresponding phosphenes (rapid blinking that made it distinguishable from the other phosphenes). Technically, we used OpenCV library to resize each video frame (320x240 pixels) to 6x10 pixels to fit the number of electrodes in the implant. The luminance of the 60 phosphenes was derived from the resized image. We also got the coordinates of the recognized faces if any. Then, the luminance of the phosphene(s) closest to the position of the detected face(s) was set to the maximum ("white phosphene"), with a 20 Hz blinking frequency. In the LOC condition, we proceeded as in the SCB+LOC condition, except that only the phosphenes corresponding to the recognized faces were displayed. The remaining phosphenes were switched off. The complete processing loop took less than 70ms, so the position of the phosphenes was updated at a minimal frequency of 15Hz.

Experiment

Subjects

Four sighted volunteers (4 men; mean age 24.5, SD 1.7; range 23-27) participated in the experiment. All were familiar with the SPV.

Procedure

Subjects had to perform a face detection task. SCB+LOC and LOC conditions were systematically assessed. We did not include a condition simulating a classical "scoreboard" rendering only because we observed that subjects were absolutely not able to detect any face at a distance exceeding 1m. The order of the two conditions was intermixed across subjects to counterbalance potential learning effects. Each subject performed 60 trials (30 trials per condition x 2 conditions). At the beginning of each session, the subject was invited to stand at a specific position in the room. Four markers were attached to his hands and shoulders in order to track their positions with the motion capture system. In each trial, 0, 1 or 2 persons were randomly placed at a specific location among 7 in the room (**Figure 3**). Then the subject had to scan his surroundings to determine the number of faces in front of him (0, 1 or 2). He was told to point his arms towards the faces (no arm if 0 face, one arm for 1 face and the two arms if 2 faces) and ask to end the trial when he was confident upon his answer. After each trial, the SPV was switched off and the next trial configuration was set up. After a block of 30 trials for a condition, the second condition was experimented. The whole experiment had an average duration of 30 minutes per subject.

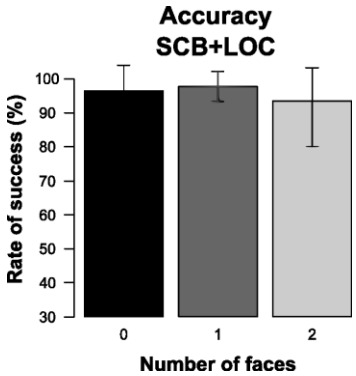

Figure 5 Accuracy per number of faces (SCB+LOC only)

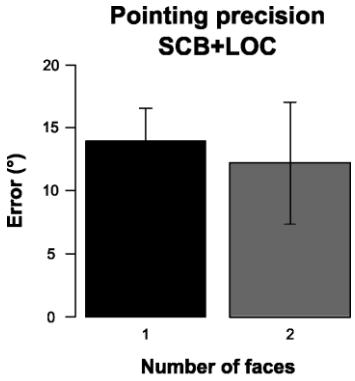

Figure 6 Pointing error per number of faces (SCB+LOC only)

Data logging

For each subject, we recorded a log file which contained all the data acquired during the experiment. The position of the markers was used to (1) determine the number of raised arms, and (2) calculate the angle in degrees between the correct and pointed directions ("pointing precision").

Preliminary results

Data analysis and statistics

We analyzed three parameters: the response accuracy (percentage of correct responses), the response time (time in seconds to give a correct answer) and the pointing precision.

We used R (R Foundation, USA) to perform the statistics. As the distribution was non-normal and the number of observations limited, we used non-parametric statistical tests. Comparisons between two groups or conditions were based on Wilcoxon tests. The significance level for all tests was set to 0.05.

Results

We first analyzed the performance in the SCB+LOC condition. All the subjects were able to perform the face detection task in this condition. The accuracy was 95.1% (SD=8.3%) and the average response time was 19.0s (SD=4.7s). The pointing precision was 13.1° (SD=3.9°). **Figure 4** presents the response time versus accuracy performance per subject.

Figure 5 and **Figure 6** show accuracy and pointing precision per number of faces in the SCB+LOC condition. The mean subjects' response time per number of faces was 16.2s (SD=4.6s) for 0 face, 18.7s (SD=2.8s) for 1 face and 22.2s (SD=5.2s) for 2 faces.

When the "scoreboard" background was switched off – LOC condition-, the accuracy (**Figure 7**, Z=-1.6, p=0.56) and pointing precision (**Figure 8**, Z=-0.56, p=0.6) were not statistically different from those in the SCB+LOC condition. The response time was lower in the LOC condition (Z=-3, p<0.001).

Discussion and future work

In this experiment, we simulated a visual implant that partially restores sight in blind people. We showed that, in spite of very low resolution, it is possible to locate small targets (here human faces) in the surroundings if the camera image is pre-processed with an object recognition algorithm. High-level information extracted from the images was used to highlight a subset of phosphenes, here corresponding to the approximate position of faces in the surroundings.

In complex, real environments, the information about faces location could be masked by the background information. We designed a second, simplified, rendering where only the phosphenes corresponding to faces location were switched on. We observed that accuracy was similar in both conditions. However the response was faster, reflecting the fact that the subjects had only one or two phosphenes to process in the absence of the "scoreboard" background.

This advantage in response time could prove useful in situations where the user wants to rapidly retrieve the location of specific objects (e.g. faces when entering a room). We suggest that a system in which the user could rapidly switch between these two modes (SCB+LOC / LOC) could increase the global efficiency of a visual neuroprosthesis.

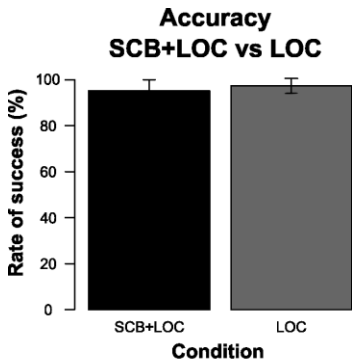

Figure 7 Accuracy per condition

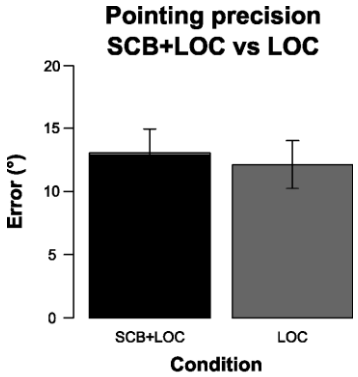

Figure 8 Pointing error per condition

Because the system relies on the extraction of high level information, we should point out limitations that typically affect object recognition algorithms. First, really small objects are difficult to detect because their apparent size rapidly shrink under the minimal detectable size. This becomes a problem for faces at a distance of 5+ meters with our system. In addition, transparent or reflective objects in general could also be very challenging to detect as their aspect depends on light conditions and surrounding objects.

Another challenge lies in the large number of models that should be created to cover all the objects that a user would like to locate. A suitable approach to recognize such a large number of objects could rely on shared databases available online [1]. Many research teams are focused on the design of these databases mainly constituted and verified though crowdsourcing [4].

To conclude, although some technical issues remain, we suggest that the preprocessing of the image in visual neuroprosthesis could subserve a great number of functions such as object recognition, text detection, navigation, etc. Interestingly, this method is suitable for low resolution implants such as those implanted nowadays.

References

1. Bigham, J.P., Jayant, C., Miller, A., White, B., and Yeh, T. VizWiz::LocateIt - enabling blind people to locate objects in their environment. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010), 65–72.

2. Campbell, P.K., Jones, K.E., Huber, R.J., Horch, K.W., and Normann, R.A. A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array. *IEEE transactions on bio-medical engineering 38*, 8 (1991), 758–68.

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

4. Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-fei, L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE (2009), 248–255.

5. Dorn, J.D., Ahuja, A.K., Caspi, A., et al. The Detection of Motion by Blind Subjects With the Epiretinal 60-Electrode (Argus II) Retinal Prosthesis. *Archives of ophthalmology*, (2012), 1–7.

6. Dramas, F., Thorpe, S.J., and Jouffrais, C. Artificial vision for the Blind: a bio-inspired algorithm for objects and obstacles detection. *International Journal of Image and Graphics 10*, 04 (2010), 531.

7. Humayun, M.S., Dorn, J.D., Da Cruz, L., et al. Interim results from the international trial of Second Sight's visual prosthesis. *Ophthalmology 119*, 4 (2012), 779–88.

8. Pérez Fornos, A., Sommerhalder, J., Da Cruz, L., et al. Temporal properties of visual perception on electrical stimulation of the retina. *Investigative ophthalmology & visual science 53*, 6 (2012), 2720–31.

9. Zrenner, E., Bartz-Schmidt, K.U., Benav, H., et al. Subretinal electronic chips allow blind patients to read letters and combine them to words. *Proceedings. Biological sciences / The Royal Society 278*, 1711 (2011), 1489–97.

10. Zrenner, E., Wilke, R., Sachs, H., et al. Patients allow recognition of letters and direction of thin stripes. *World Congress on Medical Physics and Biomedical Engineering - IFMBE Proceedings 25*, 9 (2009), 444–447.