Vision Research 111 (2015) 182–196

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Active confocal imaging for visual prostheses

Jae-Hyun Jung [a], Doron Aloni [b], Yitzhak Yitzhaky [b], Eli Peli [a,*]

[a] Schepens Eye Research Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
[b] Department of Electro-Optics Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 18 May 2014
Received in revised form 14 October 2014
Available online 3 November 2014

Keywords:
Visual prosthesis
Sensory substitution device
Retinal implant
Confocal imaging
Light-field
Clutter

ABSTRACT

There are encouraging advances in prosthetic vision for the blind, including retinal and cortical implants, and other "sensory substitution devices" that use tactile or electrical stimulation. However, they all have low resolution, limited visual field, and can display only few gray levels (limited dynamic range), severely restricting their utility. To overcome these limitations, image processing or the imaging system could emphasize objects of interest and suppress the background clutter. We propose an active confocal imaging system based on light-field technology that will enable a blind user of any visual prosthesis to efficiently scan, focus on, and "see" only an object of interest while suppressing interference from background clutter. The system captures three-dimensional scene information using a light-field sensor and displays only an in-focused plane with objects in it. After capturing a confocal image, a de-cluttering process removes the clutter based on blur difference. In preliminary experiments we verified the positive impact of confocal-based background clutter removal on recognition of objects in low resolution and limited dynamic range simulated phosphene images. Using a custom-made multiple-camera system based on light-field imaging, we confirmed that the concept of a confocal de-cluttered image can be realized effectively.

1. Introduction

An estimated 39 million people worldwide are blind (World Health Organization, 2013) and 1.2 million people in the US are legally blind and about 10% of them are functionally blind (American Foundation for the Blind, 2011). Although blind people can access text through braille and text to speech, independent mobility indoors and outside is limited and largely relies on the long cane. Blindness limits numerous activities of daily living (Brown et al., 2001; Kuyk et al., 2008), particularly tasks requiring visual search and object recognition. As a result, many pursuits (vocational and social) are limited, especially when blindness occurs in adulthood (Horowitz, 2004).

A number of implantable prosthetic vision systems have been developed (Margalit et al., 2002; Ong & Cruz, 2012). Retinal implants, such as the Argus II (Second Sight Medical Products, Sylmar, CA) (Ahuja & Behrend, 2013) and Alpha IMS (Retinal Implant AG, Kusterdingen, Germany) (Stingl et al., 2013) recently received FDA approval in the US and the CE mark in Europe, respectively. Noninvasive sensory substitution devices (SSDs) have been developed, such as the tactile graphic display (Chouvardas, Miliou, & Hatalis, 2008), BrainPort V100 (Wicab, Middleton, WI) tongue stimulation (Nau, Bach, & Fisher, 2013), and vOICe (MetaModal, Pasadena, CA) auditory vision substitution (Ward & Meijer, 2010).

Most of these systems use a video camera and convert the high resolution scene captured into a format that can be conveyed by the system transducer to the sensory organ. Although partial restoration of vision through the prostheses is expected to help improve the daily life of blind people, the utility of current visual prostheses is limited by low spatial resolution, low dynamic range (the number of displayable or perceivable gray levels), and a narrow visual field. The physical limitations of electrodes in implants and other physiological stimulators in SSDs restrict the resolution and dynamic range that can be delivered to the user. The current electrode count of the Argus II retinal implant is 60 (10×6) electrodes (Ahuja & Behrend, 2013) and expected to be about 1000 electrodes in next versions (Singer et al., 2012), and Alpha IMS has 1500 electrodes (Stingl et al., 2013). Similar limitations apply to most other SSDs. For example, the BrainPort V100 has only 400 (20×20) electrodes (Nau, Bach, & Fisher, 2013) to stimulate the user's tongue. The dynamic range of most SSDs is limited to binary (on and off) or at most 3 or 4 levels (Chouvardas, Miliou, & Hatalis, 2008). While the Argus II is capable of generating 31 brightness levels

* Corresponding author at: The Schepens Eye Research Institute, 20 Staniford Street, Boston, MA 02114-2500, USA. Fax: +1 617 912 0112.
E-mail address: eli_peli@meei.harvard.edu (E. Peli).

http://dx.doi.org/10.1016/j.visres.2014.10.023
0042-6989/

(Second Sight Medical Products Inc., 2013), only 4–12 levels of dynamic range were successfully distinguished by patients in simple just-noticeable-difference experiments (Chen et al., 2009b). In addition, the dynamic range for different visual prostheses is usually limited to less than that (Rizzo et al., 2003b) and only binary dynamic range has been used for most test and calibration (Ahuja & Behrend, 2013; da Cruz et al., 2013; Second Sight Medical Products Inc., 2013).

The visual field of retinal prostheses is on the order of 10° (Ahuja & Behrend, 2013), half the field diameter that qualifies as legal blindness, and with a visual acuity of worse than 20/1260 (Humayun et al., 2012). Mean acuity score with the BrainPort was reported as 20/5000 (Nau, Bach, & Fisher, 2013). With these limitations, reading even a short word using the Argus II requires minutes (Ahuja & Behrend, 2013) and interpreting a natural image or a scene while walking is enormously difficult (Weiland, Cho, & Humayun, 2011).

Although the performance improvements of visual prostheses are often optimistically projected to overcome technical barriers with increased electrode density (number of electrodes per degree), a real hurdle lies within the biological limitations of the interactions between the sensing organ and the stimulator that bound the likely possible resolution (Rizzo et al., 2003a, 2003b). Even if the electrode density is increased it is unlikely that visual perception will increase in proportion to the increase in density. Crosstalk between electrodes limits the improvement in effective resolution (Horsager, Greenberg, & Fine, 2010; Wilke et al., 2010), and that effect is expected to increase with higher density. The perceived dynamic range attained with each electrode varies. Even if the theoretical dynamic range from different levels of electrode stimulation exceeds 8 levels and each electrode is calibrated individually, the effective dynamic range does not increase proportionally (Chen et al., 2009b; Palanker et al., 2005; Second Sight Medical Products Inc., 2013). Until improved system interfaces are developed, improving image processing to deliver the most effective images to the stimulator is a practical and promising approach that will remain useful even when prostheses with higher effective resolution and dynamic range become available.

Visual clutter causes crowding and masking, thus reducing performance of tasks such as object segmentation, recognition, and search (Rosenholtz, Li, & Nakano, 2007). Fig. 1a illustrates typical real-world visual clutter caused by a complex background, where the near object (pedestrian) is cluttered by background objects (tree and building). While an observer can easily separate such objects for recognition in a high resolution and color image (Fig. 1b), with limited resolution and dynamic range (Figs. 1c and d) background clutter may mask bordering objects. The low resolution and dynamic range phosphene-like images created by current systems are difficult to interpret, even when the simulated images are examined with normal vision (Chen et al., 2009a; Parikh et al., 2009; Wang, Yang, & Dagnelie, 2008). Although a few studies (Humayun et al., 2012; Nau, Bach, & Fisher, 2013; Zrenner et al., 2011) have shown that letters and objects can be recognized by visual prosthesis users, the patients' performance was typically demonstrated under an ideal experimental condition, where the high contrast target object is presented in front of white or other uniform background. The reported success demonstrated in such clean laboratory settings without background clutter does not represent the visual prostheses' practical utility under real-world conditions, where a visual prosthesis with an imaging system that can effectively suppress background clutter and show only the object of interest (OI) is needed, as illustrated in Fig. 1e.

Effective compression of the camera's video to match the limited resolution and dynamic range of the prosthetic systems is crucial, but so far only basic image processing techniques have been applied (Chen et al., 2009a), such as binary thresholding (or coarse quantization in the spatial and dynamic range domains), edge detection, and image segmentation. Other higher-level analyses based on image saliency (Al-Atabany et al., 2013; Parikh, Itti, & Weiland, 2010; Weiland et al., 2012) or face detection (Li, 2013) were proposed for targeting (selecting a portion of the scene). These approaches are orthogonal to the problem we are addressing. For example, computer-vision tools may be used to segment the image into regions or even distinct (identified) OIs (e.g., faces). The segmented image can be used to present a schematic or iconic illustration, instead of an image, making it potentially more suitable to the limited capability of the prostheses. This approach was suggested for optogenetic prostheses (Al-Atabany et al., 2013), and for retinal prostheses (McCarthy, Barnes, & Lieby, 2011). In the latter case, a depth camera using structured light (Boyer & Kak, 1987) was used to help with the segmentation task. Segmenting an image is not sufficient, without some sort of additional recognition to isolate the OI and suppress the remainder.

Various types of depth cameras can be used to obtain 3D distance information that may be helpful in segmenting an OI, and such techniques have been applied to visual prostheses (Hao, Ro, & Zhigang, 2013; Li, 2013; Lieby et al., 2011; McCarthy, Barnes, & Lieby, 2011). A structured light camera (Kinect, Microsoft, Redmond, WA) or time of flight camera (Lange & Seitz, 2001) are on one end of the spectrum for acquiring 3D information, while stereo-cameras or multiple-cameras (Lieby et al., 2011; Hao, Ro, & Zhigang, 2013) are on the other. Although infrared (IR)-based tech-

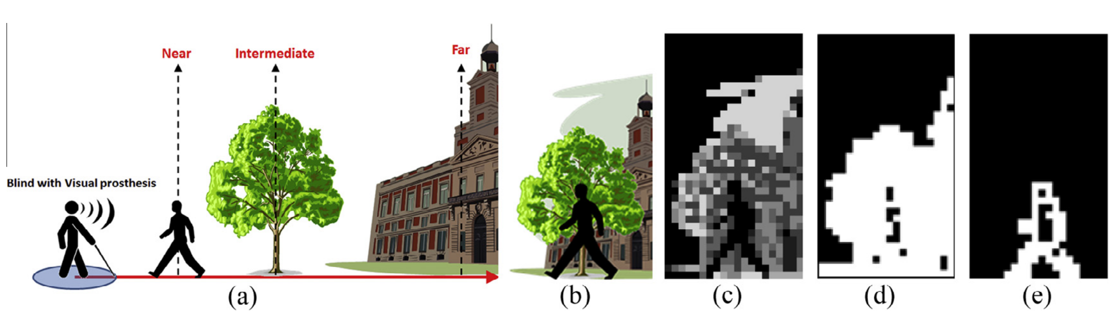

Fig. 1. Illustration of the proposed removal of background clutter for visual prostheses. (a) A blind person with visual prosthesis facing a schematic natural three-dimensional (3D) scene that includes a pedestrian in front of a tree and a building behind the tree. (b) The overlapping objects at different depths that clutter each other are captured by a head-mounted camera. In the color high resolution image, the overlapping objects of interest (OIs) can be easily separated perceptually. (c) Following image compression into low resolution (about 1,000 pixels), even with 8-bit grayscale, recognition is severely impacted. (d) Compressed binary image (simulated phosphene vision) at the same low resolution makes it difficult if not impossible to recognize the objects. (e) If the background clutter is removed by using image processing or other imaging technology, only the OI (e.g., the nearest pedestrian) will remain, thus object recognition through the visual prostheses will be improved. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)