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/© 2014 Elsevier Ltd. All rights reserved.

(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-

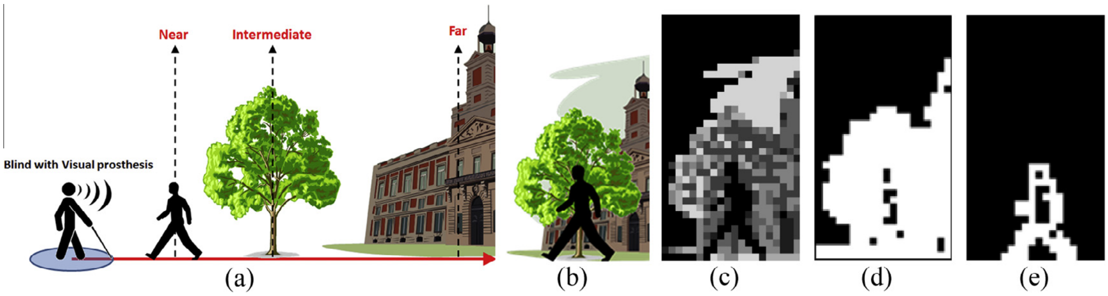


Fig. 1. Illustration of the proposed removal of background clutter for visual prostheses. (a) A blind person with visual prosthesis facing a schematic natural three-dimensional (3D) scene that includes a pedestrian in front of a tree and a building behind the tree. (b) The overlapping objects at different 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.)

nologies such as the structured light and time-of-flight cameras effectively measure the depth information, the utility of IR technology outdoors is limited by interference from the IR radiation of sunlight (Chéné et al., 2012; El-laithy, Jidong, & Yeh, 2012). Stereo or multiple cameras are not limited by outdoor use. However, correctly calculating depth from disparity is difficult and requires high computational power and time. Object segmentation and recognition algorithms are not very accurate, not easy to implement, and require high computational power. Even if the depth map is accurately extracted using an IR-based depth camera, the need remains for additional depth segmentation and object recognition processes to isolate and display only the OI and remove the background. Computer-vision tools needed for that are prone to errors around the edges of objects, exactly where we want to suppress cluttering in visual prostheses.

Most importantly, an interactive approach, allowing the user to select OIs from a small subset of depth planes is much more effective than image processing designed for machine vision. We propose an improved imaging system for visual prostheses that captures a confocal image at a depth (focal) plane selected by the user, and presents only the OIs in-focus at that depth plane. The system automatically suppresses background clutter from objects (out of focus) in other depths. The user's intent, familiarity with the environment, and situational awareness will guide real-time selection of the depth plane while searching in the depth direction. Our system also limits the search for OIs by pre-selecting depth planes where objects are likely to appear by measuring the coherence of edges at each depth with the edges appearing in a wide depth of focus image of the scene. The user can also actively scan laterally and vertically, reducing the impact of the limited field-of-view, either through a manual interface or more naturally with head movements. Then the user can actively zoom in on detected/selected objects for better detail. We call this "*active confocal de-cluttering*". In Section 2, we first show how the confocal de-cluttering process can be implemented. Section 3 assesses the benefit of de-cluttering, and Section 4 describes and demonstrates implementation with light-field technology.

2. Active confocal de-cluttering

Our proposed system of active confocal imaging for visual prostheses is composed of three stages; confocal image generation, de-cluttering, and image compression into a format suitable for visual prostheses. Confocal images from depth-sliced information are widely used in tomographic medical imaging, microscopy, and inspection, based on technologies including X-ray, CT, and MRI, confocal scanning laser ophthalmoscopy and optical coherence tomography used in retinal imaging. These methods scan and capture multiple narrow depth of field (DOF) or tomographic images with changing focal planes, which capture only objects in a focal plane, suppressing other depth planes by blur or blocking light. Similarly, a simple narrow DOF camera lens (low *f*-number) can capture a depth sliced image of OIs at a focal plane with blurred background in other depths, and generate confocal images at different depth planes. Recently, another confocal imaging method based on light-field was developed and commercialized (Harris, 2012; Ng et al., 2005) which will be discussed in Section 4. In active confocal imaging for visual prostheses, as in other applications, the main purpose of any confocal image capture method is suppression of clutter from other depths.

Fig. 2 illustrates the difference between compressed images obtained with conventional wide-DOF imaging and narrow-DOF confocal imaging. While a conventional camera image with wide DOF (Fig. 2a) focuses on both the OI and the background, the narrow-DOF confocal image (Fig. 2e) highlights only the OI (cup) at the selected depth plane against a blurred background (bookshelves). Even though, when viewed with normal vision the confocal image naturally suppresses the background clutter and emphasizes only the OI at the selected depth plane, it is insufficiently effective when applied with the high level of compression common in visual prostheses.

With the limited resolution and dynamic range of current visual prostheses, additional processing is required to suppress or remove the background clutter that is only partially suppressed by blur in the confocal image. This can be achieved if the confocal image is high-pass filtered or analyzed by some other blur metric followed by thresholding set to more completely exclude the blurred background. We name this process "*confocal de-cluttering*".

Fig. 3 shows versions of the conventional and confocal images of Fig. 2, processed via edge filtering. The edge images in Figs. 3a and c were obtained by Sobel edge detection (Sobel & Feldman, 1968) as an example, but any other edge detection methods, high-pass filter, or blur-metric algorithm (Lee et al., 2008; Park, Hong, & Lee, 2009) with appropriate thresholding can be applied to de-clutter the blurred background effectively. The confocal de-cluttered image shows only an outline of the OI and the blurred background is removed by edge filtering, whereas edges of the background in the compressed conventional image clutter the OI. Although we chose a clear object (cup) to be recognized as an example in Fig. 3, the handle of the cup in a conventional compressed edge image (Fig. 3b) is hardly recognizable. However, that detail in the compressed confocal de-cluttered image (Fig. 3d) is recognizable, with not only the shape of the cup but also the handle visible. In Section 3, a pilot study further demonstrates the positive impact of background clutter removal on recognition of compressed binary images.

3. Impact of background removal on object recognition

To determine the impact on object recognition of background de-cluttering using confocal imaging and its interaction with resolution, we conducted a preliminary study using an image dataset of objects captured by both narrow and wide DOF lens settings.

3.1. Materials

We created a dataset of images of 20 objects photographed under two conditions. Household items and office equipment were photographed in front of complex backgrounds as shown in Fig. S1 in the online supplement. Each object was captured by both narrow and wide DOF settings controlled by the *f*-number of the camera lens to simulate the confocal and conventional images, respectively. The images were captured by a NEX-3N (Sony, Tokyo, Japan) mirror-less camera with 50 mm focal length, minimum *f*/1.7 lens (Minolta, Osaka, Japan). We captured the same scene once with *f*/1.7 for the narrow DOF confocal image and once at *f*/22 to represent a conventional camera with wide DOF, as in the micro cameras used in cell phones and likely to be used in prosthetic vision devices. Although *f*/22 is not a typical setting for the indoor scene, we used it to clearly show the effect of cluttering by the background. The horizontal visual angle of the camera lens was 25°. The dataset images were taken from about 80 cm in front of the objects to maintain an angular size of the objects of about 10°.

In framing the photos and focusing the camera, we assumed that the prosthetics user would aim the camera at the OI and select the depth plane of the object (in Section 4.2, we describe a method for automatically pre-selecting a few depth planes to support efficient scanning in depth). The image viewpoint was selected to cover the whole object and emphasize the outline and distinct features of each object (the most recognizable viewpoint). We placed

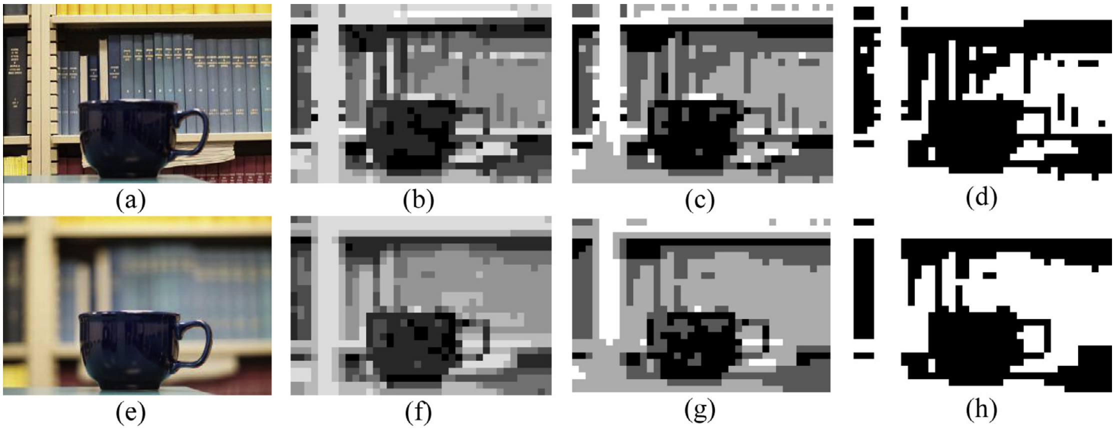


Fig. 2. Comparison of the effect of compression (low resolution and dynamic range) on a conventional (wide DOF, a–d) and confocal (narrow DOF, e–h) image; a cup in front of a complex background (bookshelves), as captured by conventional camera. When converted into low resolution (38 × 25, 950 pixels) and low dynamic range images such as (b) 8-level, (c) 4-level, and (d) binary, the background detail of the wide DOF image clutters the OI more as the dynamic range decreases. (e) With the scene captured by using confocal imaging (narrow DOF) at the selected depth plane, only the OI is in-focus and the background is naturally suppressed by blur. However, the background suppression in the compressed images (f–h) is not as apparent as in the original image. As dynamic range gets lower, the natural background suppression effect of confocal imaging is diminished.

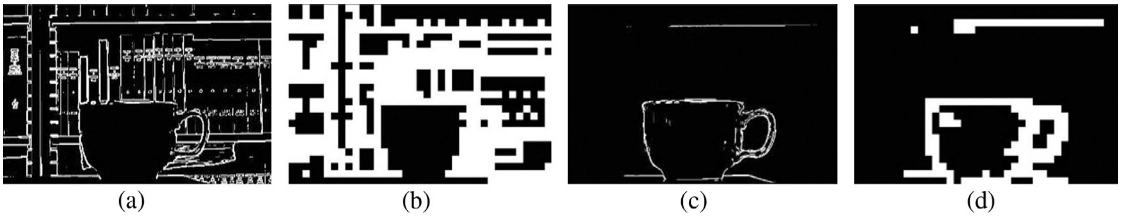


Fig. 3. Illustration of the impact of the confocal de-cluttering. The images in Fig. 2a and e were processed by edge detection and are shown here in (a) and in (c), respectively. Following compression of the image in (a) into the low resolution and dynamic range of visual prostheses, much detail of the background remains and clutters the OI in (b) and makes recognition difficult. With the confocal de-cluttered image shown in (c), the edge filtering removes the background clutter and leaves only the OI at the selected depth visible, even with compression, as shown in (d). The latter is easier to recognize, especially with regard to the handle of the cup.

each object in front of a complex background such as bookshelves, clothes, wires, and files and aimed the camera to have the OI in the bottom center of the frame and maximally overlap the OI and background, without including much floor or table surface. Although the images were staged purposely with high background complexity, we strove to create realistic scenes.

After capturing the images with blurred and focused backgrounds, we applied Sobel edge filtering (Sobel & Feldman, 1968) as the confocal de-cluttering process. Although numerous edge detection methods are available, we chose Sobel because it is widely used and easily implemented in real time. Object recognition of edge images has been shown to not be significantly related to the edge detection method (Dowling, Maeder, & Boles, 2004). The edge detection process was performed on the images after scaling down to a moderate resolution (492 × 327) that is consistent with the resolution of current light-field cameras discussed in Section 4. Although automated methods for selecting an optimal threshold for edge detection are available (Yitzhaky & Peli, 2003), we adjusted the threshold of the edge filter manually for each confocal image, aiming to fully remove the suppressed background clutter and leave only the OI in the edge filtered image. The same threshold was then applied to edge filtering for the paired conventional, wide DOF, image.

Following edge detection, the confocal de-cluttered image and the edge image (from the conventional cluttered image) of each object were compressed into 7 additional levels of resolution, using non-overlapping windows of: 2 × 2, 3 × 3, 5 × 5, 7 × 7, 13 × 13, 18 × 18, and 40 × 40 pixels, resulting in compressed images of: 246 × 164 (40,344 pixels), 164 × 109 (17,876 pixels), 98 × 65 (6370 pixels), 70 × 47 (3290 pixels), 38 × 25 (950 pixels), 27 × 18 (486 pixels), and 12 × 8 (96 pixels), respectively. The compression was performed using the following procedure: The ratio of edge pixels to non-edge pixels for all non-confocal edge images of the dataset was averaged and found to be 1/13 (7.7%). The compressed images were adjusted to maintain the same ratio by setting this ratio as the threshold for the compression at each window. If the number of edge (white) pixels in a compression window exceeded this ratio, the compressed pixel was set to white. If it fell below this ratio, the pixel was set to black. The same compression procedure and threshold were applied to both the conventional edge and confocal de-cluttered images.

Figs. 4 and 5 show the edge-filtered conventional and confocal de-cluttered images, respectively, compressed into 950 pixels. Although 950 electrodes is higher resolution than most current visual prostheses, it is still difficult to recognize the OIs with background clutter (Fig. 4). However, the OIs in the compressed confocal de-cluttered images with the same resolution (Fig. 5) are more likely recognizable than the compressed non-confocal edge images despite some residual noise, though it is by no means a trivial task.

The interaction of background removal with visual prosthesis resolution is illustrated in Figs. 6 and 7, where compressed images at 8 different resolution levels are compared with and without background clutter, respectively. Even at a resolution over 3,000 pixels, 2 or 3 times higher than the current or anticipated next-generation retinal implant, the complex background clutters the OI and makes the OI difficult to recognize (Fig. 6d). The compressed confocal de-cluttered image emphasizes the OI and enables recognition at a lower resolution level (Fig. 7). By increasing the resolu-

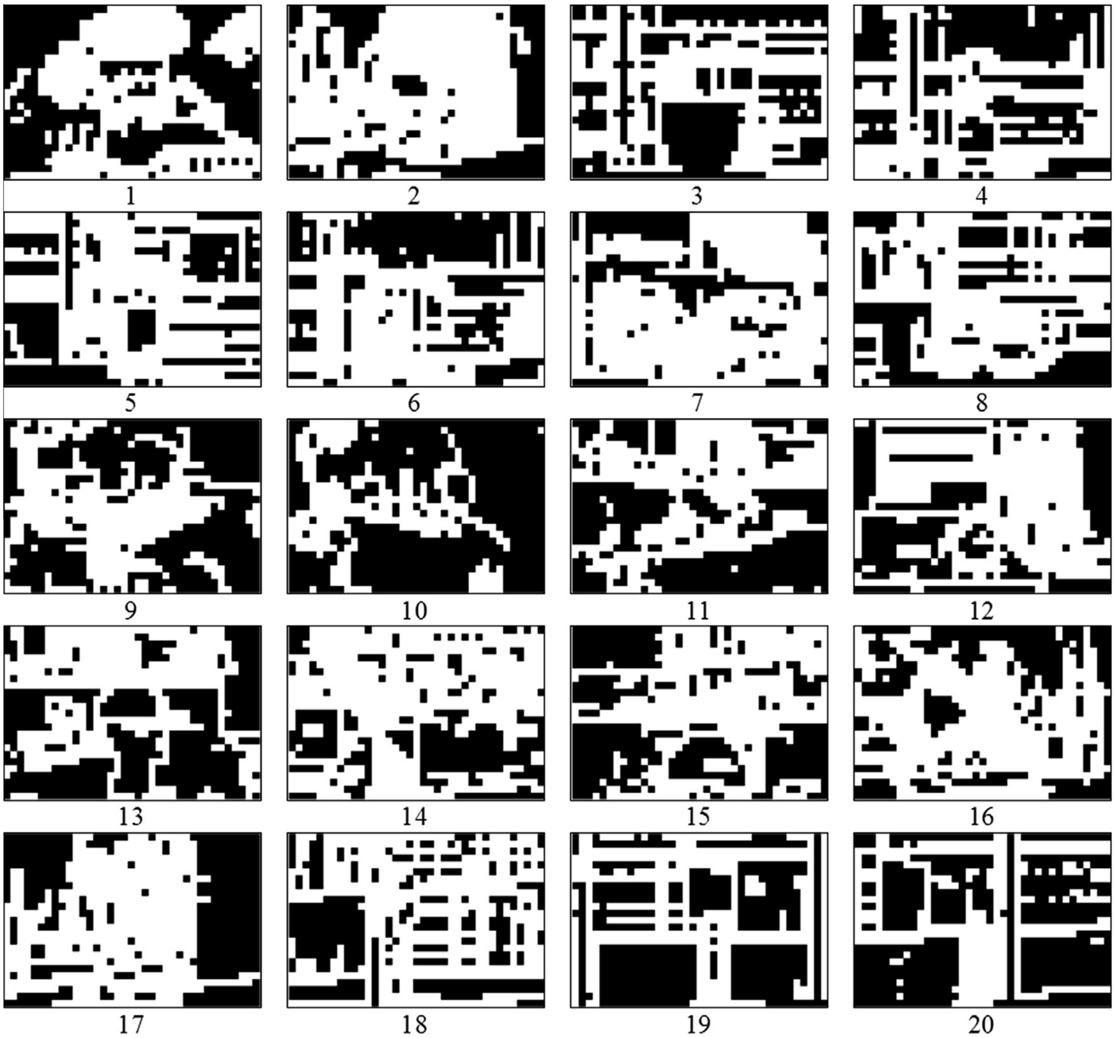


Fig. 4. The 20 dataset images in non-confocal conventional photography compressed into 950 pixels (38 by 25) following the edge detection process. Compressing the edge images results in cluttering of objects and disruption of the borders between the OI and background. To recognize the OI with these imaging, higher resolution or dynamic range is required.

tion, the recognition of the OI and its details becomes easier. In Section 3.2, we measure the object recognition rate in background de-cluttered and cluttered conditions, using this created dataset.

3.2. Object recognition test

3.2.1. Methods

A preliminary object recognition test was performed with 6 normally sighted subjects (3 women), aged 24–42, using the image dataset (Section 3.1). The study was approved by the Human Studies Committee of the Massachusetts Eye and Ear and written informed consent was obtained. The 20-object images were randomly ordered within blocks of the same compression level and same background condition (cluttered versus de-cluttered). The presentation of a block of images always started from low resolution and proceeded to higher resolution. At each compression level, randomly ordered presentations of the background-cluttered images of 20 objects were followed by a block of background-de-cluttered images. This sequence of 320 images was displayed on a 21″ P1130 Trinitron CRT monitor (Dell Inc., Round Rock, TX) at 1280 × 1024 resolution and observed by subjects from 80 cm away. The size of all images was 14.6 cm by 9.7 cm, spanning a visual angle of 10.4° × 6.9°. The image sequence was displayed at the center of the screen surrounded by a blue border so that subjects easily distinguished the area of the image.

We explained the task to subjects during a training session. First, at full resolution (160,884 pixels, 492 × 327), conventional and confocal images were shown followed by the edge-filtered and compressed (decreased resolution) images. The subjects were informed of the categories of objects presented (household items and office equipment), the average size of objects (all objects were smaller than the 21″ monitor screen), and the position of objects in the images (bottom center). The viewpoint for image acquisition was disclosed to the subjects. The specific object recognition task was then performed with background cluttered and de-cluttered images with the 8 different levels of resolution as a training session, to familiarize the subjects with the interpretation needed by the low resolution edge images. We also discussed with the subjects the nature of the edge images and the cluttering at low resolution. Following a training session with 3 different objects, subjects commenced the main task. Subjects could guess the OI or pass on difficult scenes after 1 min. If they could not name the recognized object, they were allowed to describe the use of the object, details of its shape, or specific features. The operator wrote down the subjects' responses, but no feedback or correction was provided. In deciding the veracity of responses, describing the use of the OI had a higher value than a general description of the

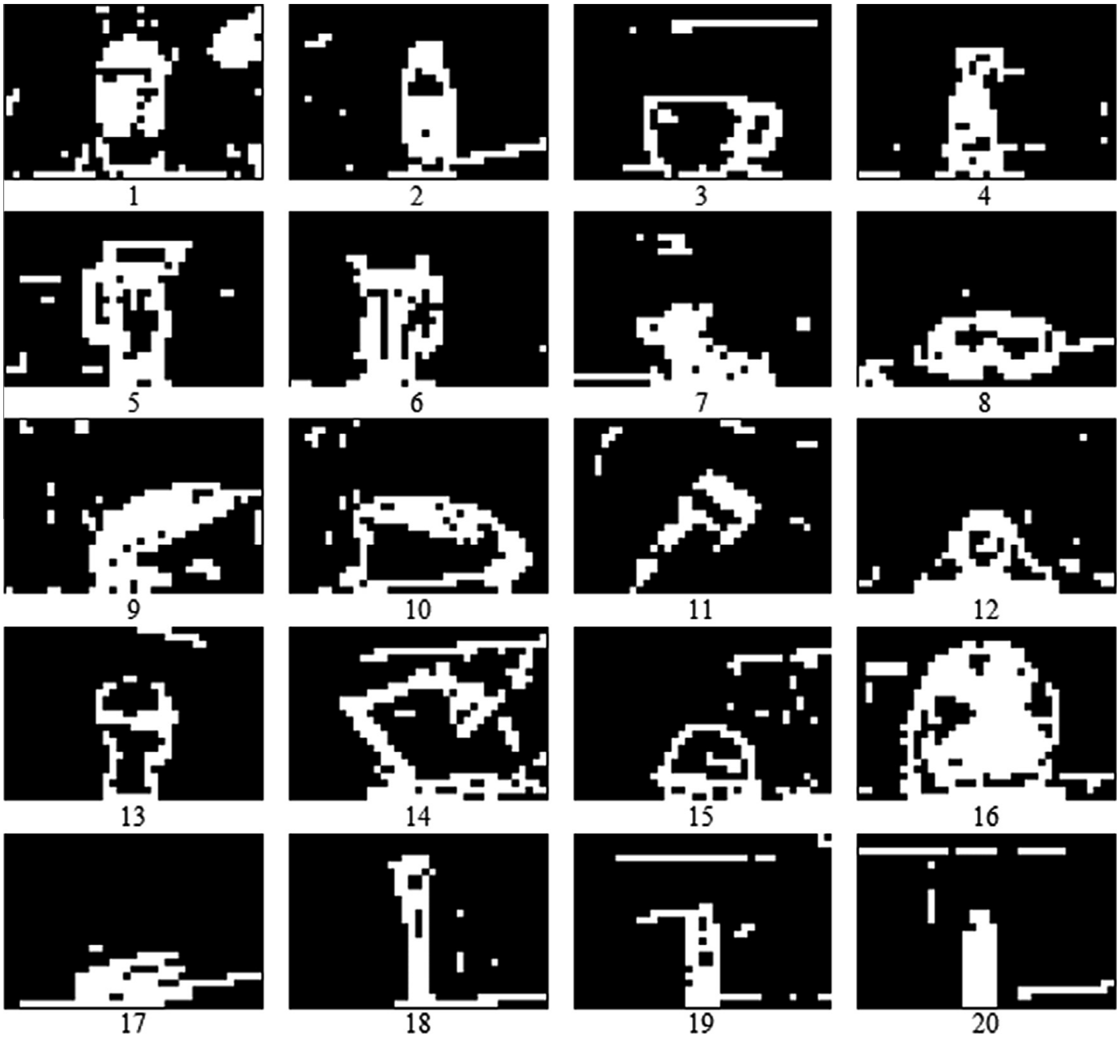


Fig. 5. The confocal de-cluttered images shown in Fig. 4 compressed in the same way. With the removal of background clutter using confocal de-cluttering, it is possible for at least a few objects to be recognized, even at this resolution.

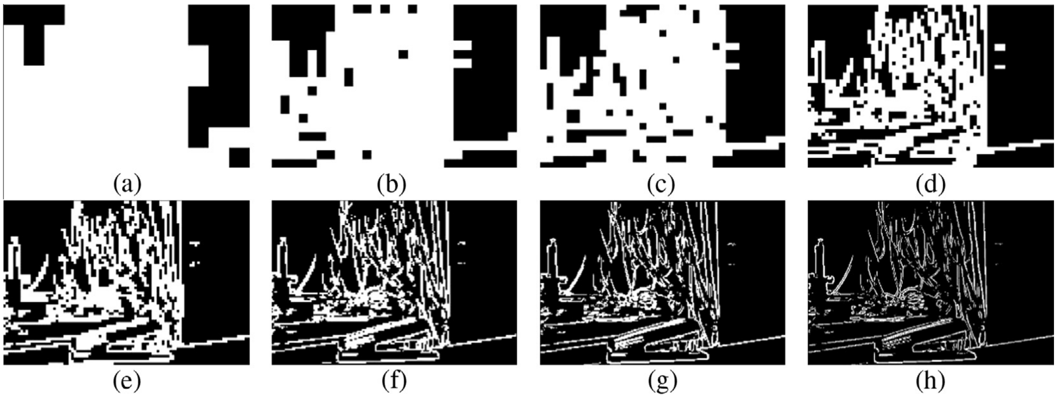


Fig. 6. A background-cluttered image compressed into different resolutions. (a) 96 pixels, (b) 486 pixels, (c) 950 pixels, (d) 3,290 pixels, (e) 6,370 pixels, (f) 17,876 pixels, (g) 40,344 pixels, and (h) 160,884 pixels. As the resolution increases, the cluttering declines and overlapped outlines are separated. However, the recognition of the OI is still not easy at least until the level shown in (e).

object shape. For example, although the shape of the helmet is much closer to an upside-down bowl than a hat or cap, we chose hat or cap as a correct response and rejected the upside-down bowl as a mere description of the shape and incorrect recognition. Because these cases were very rare (11/1,920 = 0.6%) in this experiment, these decisions did not affect the results.

We used binary logistic regression (Bender & Grouven, 1998) to estimate the impact of background removal on object recognition, to provide statistical inference of the differences among the 20 objects, and the impact of 8 resolution levels. Binary logistic regression is used to predict a categorical (usually dichotomous) variable from a set of predictor variables. With a categorical depen-

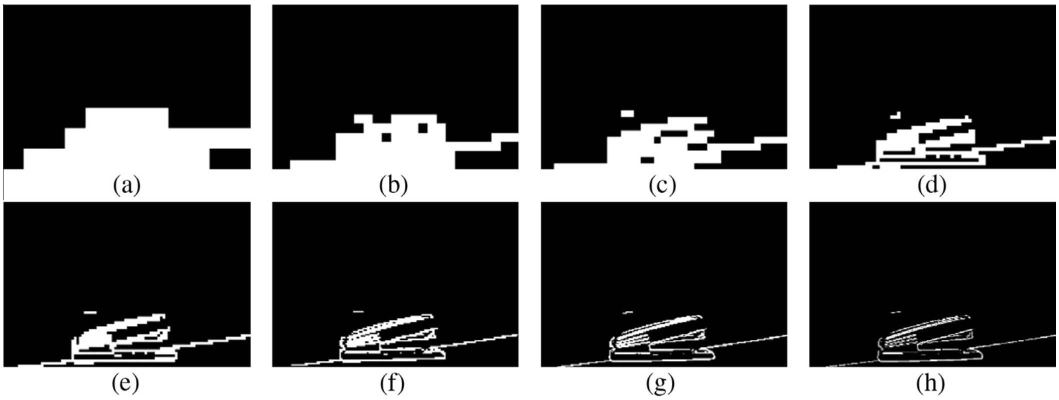


Fig. 7. The background de-cluttered image compressed into the same resolution levels as in Fig. 6. Overall, the object is easier to locate and recognize in these images than in those shown in Fig. 6. Although the background clutter is removed at all levels, details of this OI are not easily resolved below level (d). Note that zooming in on the object will improve the resolution and enable recognition at higher levels of compression.

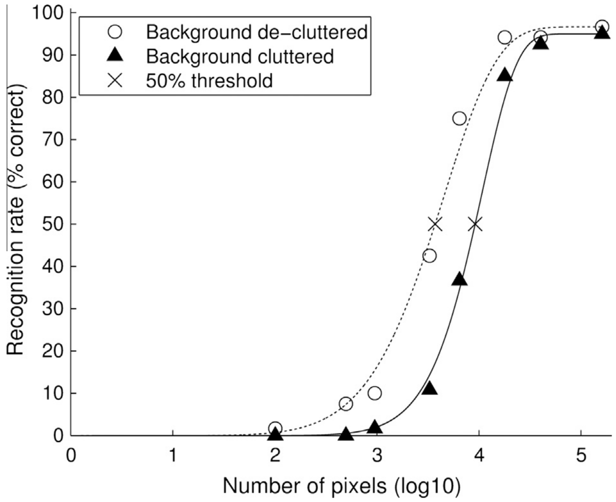


Fig. 8. The recognition rates of the 20 objects by the 6 subjects as a function of resolution. The recognition rates started to increase rapidly at about 1,000 (10^3) and about 3,100 ($10^{3.5}$) pixels in background de-cluttered and cluttered conditions, respectively. The recognition rate with the background de-cluttered condition was higher than with the background cluttered condition. Weibull psychometric functions were fitted to the data.

dent variable, binary logistic regression is often chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not normally distributed. Binary logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease (Press & Wilson, 1978). In this study, we used binary logistic regression because the response of subjects is binary (recognized or not) and other variables were a mix of continuous (log of resolution) and categorical (object number) variables.

3.2.2. Results

Fig. 8 shows the recognition rate over all 6 subjects' responses. The 20-object recognition rates of all 6 subjects (120 responses at each resolution level) are represented separately for the background cluttered and de-cluttered conditions, and are fitted with a Weibull psychometric function (Wichmann & Hill, 2001) using Psychtoolbox (Psychtoolbox-3; www.psychtoolbox.org) with MATLAB (MathWorks, Natick, MA). Because some subjects failed to recognize some objects, even at the highest resolution with background cluttered or de-cluttered conditions, the psychometric functions are not forced to reach 100% recognition. The data and fitting for individual subjects are provided in the online supplement (Fig. S2).

The recognition rate with the background cluttered condition improves from about 1,000 pixels. It saturated at about 10,000 pixels. The fitted psychometric curve for the confocal de-cluttered condition is to the left of the curve obtained with the background cluttered. The 50% recognition threshold for the conventional compressed edge images required a resolution of 8,695 (about 114 × 76) pixels, while for the de-cluttered images the same 50% performance was achievable at a resolution of 3,532 pixels (about 73 × 48). Fig. 9 shows the resolution required for a 50% recognition rate for each subject under background cluttered and de-cluttered conditions. When the compressed resolution was higher than 31,000 ($10^{4.5}$) pixels, subjects could recognize most objects regardless of the background condition. For resolutions lower than 100 pixels, most objects could not be recognized by subjects regardless of background condition. With the 1,000 to 10,000 pixel resolutions targeted by foreseeable future retinal implants, the recognition rates were clearly improved by de-cluttering (see the full curves for all 6 subjects in Fig. S2 in the online supplement).

The binary logistic regression was performed with all 1920 trials from the 6 subjects. The model and parameters were estimated in SPSS 11.5.0 (SPSS, Chicago, IL). Overall about half of the responses (1,028/1,920 = 53.5%) did not recognize and 46.5% (892/1,920) responses correctly recognized the OI. The predictor

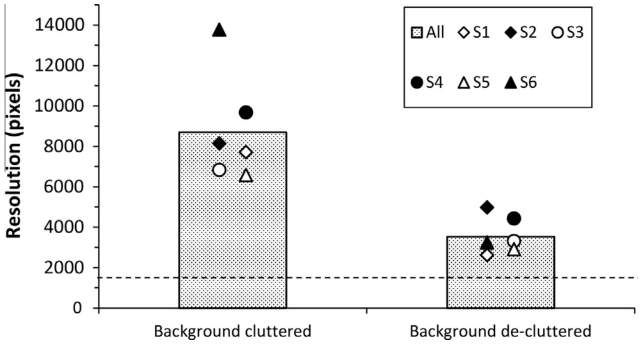


Fig. 9. The number of pixels required for 50% recognition rate by each subject under background cluttered and de-cluttered conditions. Each marker is slightly off center to prevent overlapping of markers. The 50% threshold of recognition rate over all subjects' responses is at a resolution of 8,695 pixels with cluttered background and 3,532 pixels with de-cluttered background as illustrated in gray bars. The dashed line (at 1,500 pixels) indicates the resolution of current and next-generation visual prostheses.

variables were: the background, 19 dummy variables coding the objects, and a continuous variable for resolution. The background-cluttered condition and the first object (object 1) were set as the reference for dummy variable coding. The model correctly classified 90.3% of the correct OI recognitions and 91.0% of incorrect recognitions (including no response/passes). Employing an $\alpha < 0.05$ criterion for statistical significance, the background condition, the resolution, and 14 of the object dataset dummy variables had significant partial effects.

In the binary logistic regression model from all subjects' responses, the impact of each variable was predicted by the model as an odds ratio (OR).[1] The model-predicted odds ratio for recognition of images with the background cluttered condition to the background de-cluttered condition is 5.6 ($\alpha < 0.01$), indicating that when holding all other variables constant, a confocal de-cluttered image is 5.6 times more likely to be recognized than a conventional background-cluttered image. If the recognition rate (p') of background-cluttered images is selected as reference, the recognition rate of background de-cluttered images (p) could be predicted by the odds ratio from the model as shown in the following equation:

$$p = \frac{O_R \cdot O'}{1 + O_R \cdot O'} = \frac{O_R \cdot \left(\frac{p'}{1-p'}\right)}{1 + O_R \cdot \left(\frac{p'}{1-p'}\right)}. \tag{1}$$

For example, if the conventional compressed image (background cluttered) is the reference and its recognition rate is 50%, the recognition rate of confocal de-cluttered images (background de-cluttered) is expected to be 84.9% based on the odds ratio (5.6) from the model. On the other hand, if the recognition rate of confocal de-cluttered images is 50%, the recognition rate of conventional compressed images is only 15.1%. These results are consistent with the psychometric function fitting for each subject (Fig. S2) and responses accumulated over all subjects as shown in Fig. 8. The 95% confidence interval (CI) of the odds ratio for background is from 3.9 to 8.1 verifying that background removal using confocal de-cluttering improves the object recognition performance substantially and significantly.

Similarly, the odds ratio for resolution was analyzed using the binary logistic regression model. Because the resolution was increased logarithmically, the odds ratio was analyzed by the common logarithm of the resolution. The recognition rate increases exponentially with increased resolution. The model predicted odds ratio of the common logarithm of resolution is 74.5 ($\alpha < 0.01$) and the 95% CI is from 48.9 to 113.5. This means that the recognition rate is increased 74.5 times by a resolution increment of 10 times. For example, if the recognition rate in most compressed images (96 pixels) is assumed to be 1%, the recognition rates in each resolution log step are predicted by the model to be 10.3%, 23.8%, 66.6%, 84.2%, 96.1%, 98.8%, and 99.8%.

We tried to adjust the difficulty of images in our dataset to be as uniform as possible using a similar camera viewpoint, object position, size, and background complexity. The overall difficulty of the task is highly dependent on the subject's prior experience and abilities. The relative difficulty among objects in the dataset compared with object 1 as a reference (arbitrarily selected) was analyzed by the odds ratio from the model. The relative difficulty of object recognition was analyzed using the false recognition rate because the difficulty is based on the probability of failure to recognize. Although this analysis is limited by the small sample size and generating a final data set was not the main purpose of this experiment, at least the result of this analysis showed that recognition difficulties were moderately balanced a cross samples in the data set. The detailed results are provided in the online supplement (Fig. S3 in the online supplemental materials).

We verified the impact on object recognition of background de-cluttering using a conventional camera with narrow-DOF lens. In Section 4, we present a confocal image generation method based on light-field imaging (Harris, 2012; Ng et al., 2005) and illustrate the blur-based de-cluttering process applied to the confocal image obtained from the light-field image. Then, a zooming function is included to improve object detail.

4. Active confocal de-cluttering using a light-field camera

4.1. Confocal de-cluttering based on light-field

The simplest way to acquire a confocal image without a complex optical setup is to use a low *f*-number camera lens, as we used in the preliminary object recognition test. However, the low *f*-number means the aperture size has to be wide and the focal length has to be short, which results in heavy weight and a bulky lens. The motorized mechanical and optical parts for changing focal distance also increase the volume and weight of the lens, making it inappropriate to use in a miniaturized head- or glasses-mounted camera for visual prostheses. If a scene has multiple OIs at different distances, a conventional camera lens has to change the focus for each OI and scan the whole depth range mechanically. Since this process requires mechanical adjusting of the focal distance, this method cannot be implemented for practical use, as image acquisition at high frame rates is required. Most importantly, the confocal functionality (narrower DOF) of a low *f*-number lens is not sufficient to clearly de-clutter background, because the DOF of conventional camera widens too rapidly with increasing the focal distance (Maiello, 2013).

We propose a confocal imaging system for visual prostheses based on a confocal technology called "light-field" imaging (Harris, 2012; Ng et al., 2005) to achieve confocal images effectively in a portable and cosmetically acceptable size. Confocal imaging based on light-field technology was first proposed as a 3D display technology termed integral imaging (Jung et al., 2012; Kim et al., 2014; Lippmann, 1908). A light-field image (or *elemental* image) contains all angular and spatial information of light from a 3D scene, captured by a two-dimensional microlens array (or an array of multiple cameras), where each lenslet provides a slightly different viewpoint (Harris, 2012; Ng et al., 2005). Current commercial light-field cameras such as those from Lytro (Lytro Inc., Mountain View, CA) and Raytrix (Raytrix GmBH, Kiel, Germany) use a microlens array with relay optics, where each lenslet in the micro lens array and the corresponding subset of CCD sensor pixels under it acts like an individual camera system.

Fig. 10 illustrates a simulated elemental image of the schematic 3D scene of Fig. 1. We simulated the three different plane images of the pedestrian, tree, and building to be respectively located at 1 m, 4 m, and 9 m and minified 40 times by relay optics in front of the lens array (1 mm pitch with 3 mm focal length). The angular and spatial information of the three plane images at different depths was computationally projected on the elemental image (Fig. 10 left) at the focal plane of lens array by ray tracing through each lenslet (Min et al., 2001). Each subset image of the elemental image captured by each lenslet is the same as the image that would be captured by a camera in an array of multiple cameras, but it has a reduced two-dimensional resolution (10 by 10 in Fig. 10) because of the divided CCD resolution. The ensemble captures the depth information as a trade-off for the resolution loss (inset of

[1] *OR* is defined as the ratio of the odds (O) with a variable to the odds (O') of a reference, where p is the probability of the binary event, the complementary probability is $1 - p$, and the odds of an event are defined as $p/(1 - p)$. Therefore, where p is the recognition rate with a variable and p' is the recognition rate of the reference for this variable in this experiment, the odds of reference (O') is defined as $p'/(1 - p')$.

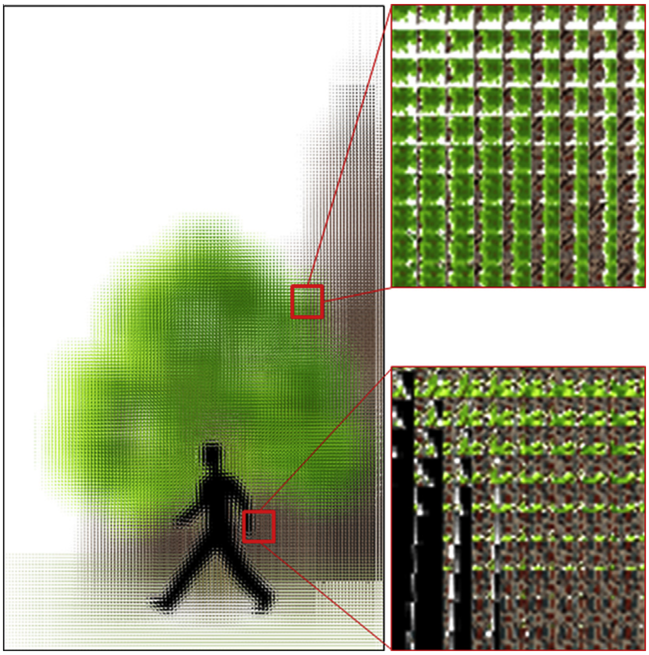


Fig. 10. Details of a simulated elemental image (light-field information) shown in two magnified insets. The simulated scene of Fig. 1 was captured by a simulated (computed) light-field camera composed of a 1 mm pitch lens array behind relay optics and in front of a CCD sensor. Each inset shows a magnified 9 × 10 subset of the elemental image. Each subset represents a different perspective view (with low resolution of 10 × 10 in this simulation) captured by a lenslet in a different position. The total light-field image contains the full 3D information of the scene.

Fig. 10). Luckily, losing some resolution in the image capture is a very low cost to pay in our application, as the image resolution needs to be compressed even further to be presented in a prosthetic vision device.

This over-informative data permits 3D visualization for display (Jung et al., 2012; Kim, Hong, & Lee, 2010; Kim et al., 2014; Lippmann, 1908), reconstruction of objects (Jung et al., 2010; Kim et al., 2013) or generation of a confocal image (Hong, Jang, & Javidi, 2004; Stern & Javidi, 2006). A 3D point captured by each lens in the array which covers the 3D point within its viewing angle can be reconstructed by a similar setup of lens array and other relay optics (Kim, Hong, & Lee, 2010; Kim et al., 2014; Lippmann, 1908). If a two-dimensional (2D) screen is placed at the depth of the reconstructed 3D point, the 3D point is projected in-focus on the 2D screen. If the screen is moved away from the depth of the 3D point to another location, an out-of-focus blurred image of that 3D point will be cast on the screen (Hong, Jang, & Javidi, 2004; Stern & Javidi, 2006). Thus, only the 3D points located at the depth of the screen are focused and other points at each depth plane are blurred.

If the reconstruction and projection processes are performed computationally using ray tracing (Hong, Jang, & Javidi, 2004; Stern & Javidi, 2006) rather than the optical reconstruction with a lens array, each projected image on the screen at the different depths is a confocal image and the depth plane of the screen is the confocal distance. It does not require an additional depth map extraction process (Hong, Jang, & Javidi, 2004; Stern & Javidi, 2006) and can also generate all-in-focus image (cluttered image) in addition to the confocal image (Harris, 2012; Ng et al., 2005). Because light-field confocal imaging is based on the computational projection of a subset of the elemental image through each lens in the array and the summation of their brightness, each simple projection calculation can be performed in parallel and the computational load needed to sum the projected elemental image pixels is low enough to be realized in real-time (Harris, 2012; Kim, Erdenebat, et al., 2013; Kim et al., 2013). Fig. 11 shows the computationally-generated confocal images in different depths from the simulated elemental image of Fig. 10.

Whereas optical confocal imaging as used in microscopy systems captures only one distance (confocal image) per frame, a light-field camera can capture the elemental image in one exposure/frame and generate multiple confocal images through rapid computation, without any mechanical movements. In addition, the light-field camera can generate the confocal image with a DOF narrower than a single narrow-DOF lens. Whereas the DOF in a single lens is limited by the designed *f*-number (focal length divided by aperture size), the *f*-number of a light-field camera can be controlled by the synthesized aperture of the light-field camera (Levoy et al., 2004; Ng et al., 2005) instead of the physical aperture of a single camera lens. With a multiple-camera array, the same lens can create a much smaller *f*-number using a synthetic aperture (Levoy et al., 2004; Ng et al., 2005) determined by the distance between cameras.

Following generation of the confocal image an additional decluttering process is needed to remove/suppress blurred background clutter. Various methods may be implemented to suppress the low spatial frequency portion of blurred background clutter for light-field confocal imaging, such as edge detection (Aloni & Yitzhaky, 2014) and blur metric (Lee et al., 2008; Park, Hong, & Lee, 2009). Fig. 12 shows the confocal de-cluttered versions of the images in Fig. 11 using Sobel edge detection.

Fig. 13 shows the final result of the confocal de-cluttered images, compressed into the low resolution (980 pixels) and dynamic range (binary) format of prosthetic vision devices and SSDs. Although object recognition from these compressed images is challenging, the user can gain some situational awareness by scanning through the depth planes shown. An automated technique that eliminates the non-object planes (e.g. Figs. 13b and d) is discussed in Section 4.2.

Once an OI is found in an object-containing plane, the user may want to zoom in on the detected/selected OI for more details. The zooming can be manually controlled and used to fill the field of view of the prosthesis or even overfill the field of view, and be used in conjunction with horizontal and vertical scanning which is very natural and easy to conduct (Hwang, Peli, & Peli, submitted for publication). Zooming in this case may not necessarily involve any magnification, scaling, or mechanical/optical movement; instead the high-resolution confocal de-cluttered image can be cropped and then compressed to a lesser extent to fit the dimensions of the prosthesis. Because the confocal de-cluttered image has higher resolution than the resolution of visual prostheses, the compressed confocal de-cluttered image (Fig. 14c) of the cropped image (Fig. 14b) includes much more OI detail than the original fully-compressed image (Fig. 13a). The impact of the zooming by cropping is not coming from magnification per se but rather from the lower level of compression applied. Note that the zooming by cropping is also more economical computationally.

4.2. Automatic detection of confocal distance of objects of interest

The user of this system can scan in depth and identify planes with a potential OI by changing the confocal depth plane. This requires scanning through the whole depth range, which may be inefficient. To reduce the scanning time, we employed a new method that isolates the focused regions from the reconstructed image planes, enabling automatic detection of planes that may contain OIs (Aloni & Yitzhaky, 2014). This algorithm is based on the assumption that the object details located at the depth of a reconstructed plane are fully focused, while objects at other depths are blurred. The focused regions in the reconstructed plane images

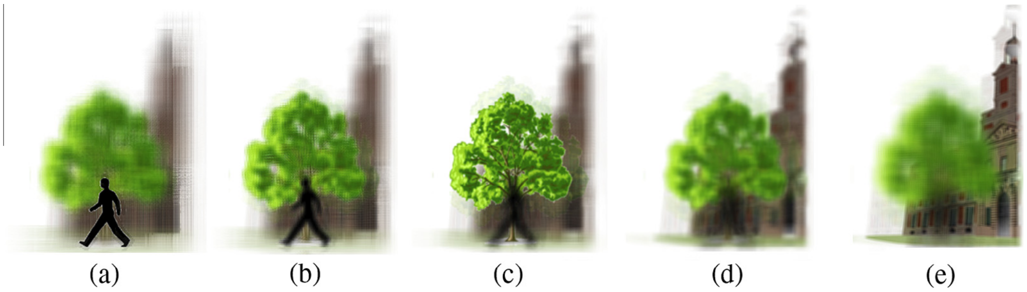


Fig. 11. Confocal images (308 × 385) in different depth planes generated from a simulated elemental image frame obtained computationally (Fig. 10) from the simulated 3 plane scene of Fig. 1. (a) The confocal images at the depth plane of the pedestrian (1 m), (b) between the pedestrian and the tree (2.5 m), (c) of the tree (4 m), (d) between the tree and the building (6.5 m), and (e) of the building (9 m). Animation 1 in the online supplement shows the confocal image sequence being scanned between near and far in depth.

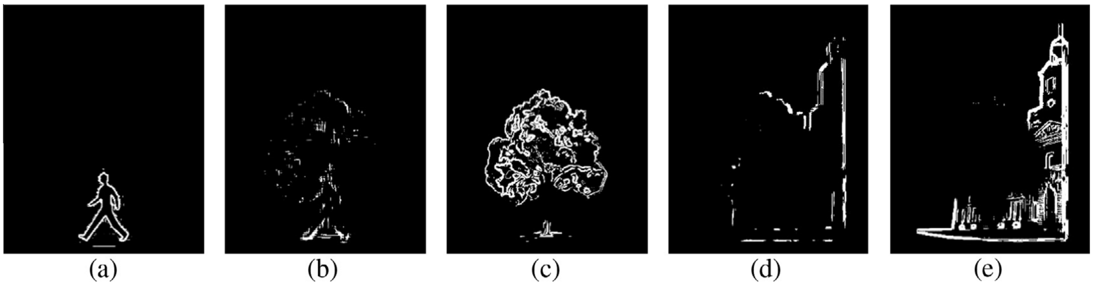


Fig. 12. Confocal de-cluttered images (308 × 385) at the different depth planes shown in Fig. 11, achieved through Sobel edge detection. Note that although there are only 3 objects in different planes in the original simulated scene, additional depth planes between objects were selected (in b and d). These intermediate depth planes (b and d) do not provide as good a result as the confocal de-cluttered image at object planes (a, c, and e). Animation 2 in the supplement shows the confocal de-cluttered image depth sequence obtained from one elemental image frame.

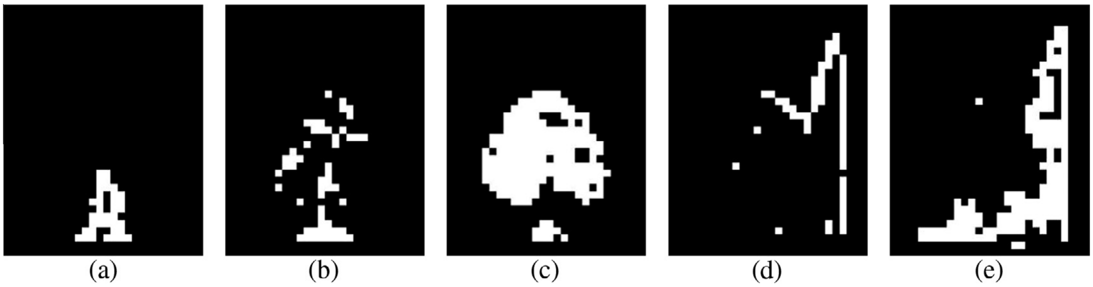


Fig. 13. Confocal de-cluttered images of Fig 12 are compressed to fit the limited resolution of a 980 pixel (28 × 35) visual prosthesis. Animation 3 in the supplement shows the compressed confocal de-cluttered images in sequence, obtained from one elemental image frame.

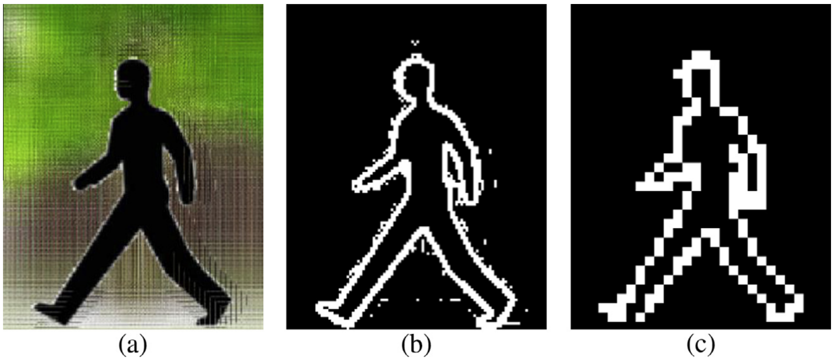


Fig. 14. Effect of zooming using cropping of the high resolution confocal image before confocal de-cluttering and compression. (a) Zoomed OI in the high resolution confocal image of Fig. 11a using cropping and therefore requiring a lesser compression. (b) The confocal de-cluttered zoomed image has a higher level of details. (c) With zoom preceding compression, more detail can be preserved in the low resolution compressed image than the compressed result without zooming of Fig. 13a.

consist of higher spatial frequencies compared to the blurry regions. We enhance the sharp edge regions using a gradient-like operation (in three directions) obtained using the first-scale Haar wavelet transform (Mallat, 1989). Then, with an adaptive threshold in each sub-band of the wavelet we detect the sharpest edge locations (Aloni & Yitzhaky, 2014). As the threshold applied to the wavelet sub-band is decreased, the number of detected edge pixels is increased. In the adaptive process, the threshold is adjusted to set the number of detected edge pixels to 0.5% of all pixels in the sub-band (Aloni & Yitzhaky, 2014).

To determine the depth of objects in the light-field image, first, the edge detection operation is repeated for reconstructed planes at many distances and also for a center view image (center subset of elemental image-wide DOF). Then, for each distance, the number of edge pixels in each confocal plane that overlap (collocate) with edge pixels of the the center view image (all-in-focus image or center elemental image) is counted. The rate of overlapping edge pixels is expected to achieve local maxima at depth planes that contain objects, because objects at these planes appear sharp in both the reconstructed planes and the center view image, thus producing edges at similar locations. Edge pixels at blurred planes are either suppressed or shifted slightly and thus do not overlap with the sharp edges in the elemental image, resulting in a smaller number of intersecting edge pixels for these planes. A graph showing the result of this process is presented in Fig. 15, as calculated for the image shown in Fig. 16. Two local maxima are seen; one very sharp at about 0.6 m and one less distinct but clear at about 3 m.

The image and the corresponding edges reconstructed from 0.6 m distance are presented in Fig. 16d–f. The confocal image (Fig. 16d) shows two objects in focus (the mug and the camera) at about the location of the reconstructed plane at 0.6 m. The confocal de-cluttered image using edge detection is shown in Fig. 16e. In this image, only edges of these two objects are detected, while edges of objects at other depth are removed. The compressed 30 × 26 pixels edge-image version is shown in Fig. 16f. Compared to the compressed background cluttered image shown in Fig. 16c, the features of the two objects in Fig. 16f better represent the objects, while in Fig. 16c these features are largely masked by the edges of the background.

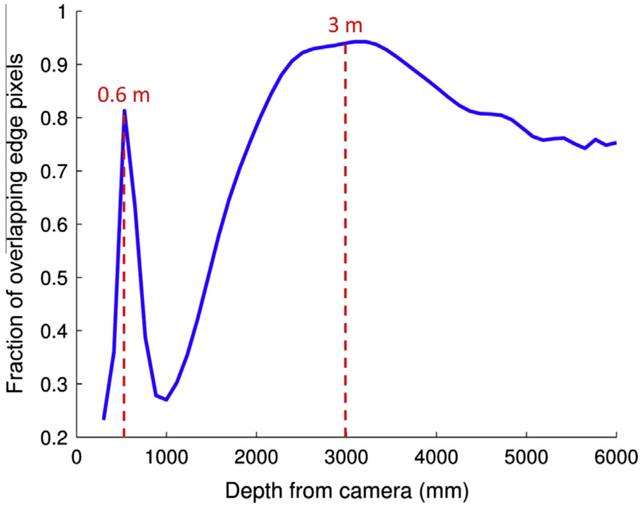


Fig. 15. Estimation of object depth planes. The fraction of overlapping (collocating) edge pixels between the edges of the center view image and the edges in 200 confocal images reconstructed at steps of 30 mm apart. The first maximum at 0.6 m distance from the camera indicates the location of the OIs in front (mainly the camera and the mug in Fig. 16a). The next maximum is around 3 m, which is the distance to the background.

Figs. 16g–i present results for a confocal image at 3 m distance from the camera, which is roughly the distance of most of the background objects. In the confocal image shown in Fig. 16g the two objects in the foreground appear blurred. These objects disappear in the edge image (Fig. 16h) and consequently in the compressed version (Fig. 16i). It can be seen that in this case the computer screens at the rear are somewhat more recognizable than in Fig. 16c where they are cluttered at the bottom by the foreground objects.

5. Discussion

We propose confocal imaging to suppress the background clutter that impedes the recognition of OIs, especially when presented in the limited resolution and dynamic range typical of current or anticipated visual prostheses. These problems are evident in simulation images shown by others (e.g., Zhao et al., 2010). In a preliminary study, we found that a confocal de-cluttered image enabled better recognition than a background cluttered image when compressed similarly. We illustrated the feasibility of obtaining confocal images via light-field cameras and the utility of the light-field data they generate. We also propose that the system could be active, where the user controls the parameters applied at various instances. The active nature of the proposed system is designed to benefit from the situational awareness of the user in general and particularly in selecting the confocal plane to be examined. The latter aspect has not been addressed in the current paper experimentally but is an important component of the proposed system.

For a variety of reasons, we used real objects in a recognition task, rather than the more commonly used multiple choice tasks, such as visual acuity, contrast sensitivity, or the discrimination between a few objects. First we argue that crowding and possibly masking by background clutter are applicable and relevant mostly to natural object recognition in a natural environment, though clearly letters can be crowded as can the direction of Gabor patches. However, the nature of visual acuity, contrast sensitivity, and object discrimination testing as performed with vision prostheses renders the stimuli free of the crowding effect (Humayun et al., 2012; Nau, Bach, & Fisher, 2013; Zrenner et al., 2011), which is the focus of our approach and proposed solution. Second, we argue that multiple choice testing, while a perfectly good method for measuring the threshold performance of the human or animal visual systems, is not sufficient to prove that prosthetic vision can deliver object recognition. Humans are excellent at pattern discrimination, and thus can learn to discriminate multiple choice targets without being able to recognize them. The ability to discriminate contrast or even orientation of Gabor patches with a prosthetic vision system does not assure an ability to transfer that capability to visual perception of objects. Observers can learn to use sounds to discriminate spatial patterns coded in some way. Yet there is little confidence that such performance will lead to auditory recognition of complex visual objects (not withstanding the claims for auditory prosthetics, Ward & Meijer, 2010). We argue that object recognition testing (not multiple choice testing of object discrimination) is crucial to evaluation of prosthetic vision. Many results demonstrated with prosthetic vision systems, especially mobility related tests, clearly show the user scanning the narrow camera field of view back and forth using head movements across high contrast markers (Cooke, 2006; Mcnamara, 2007; Neve, 2011). This operational mode is similar to the operation of radar. The radar indeed functions very well in detecting small targets in empty non-cluttered scene, such as the sky or the sea, but is not useful in the terrestrial environment. Thus we believe that the problem of clutter will be a significant impediment to the use of these devices for mobility. More relevant to our dis-

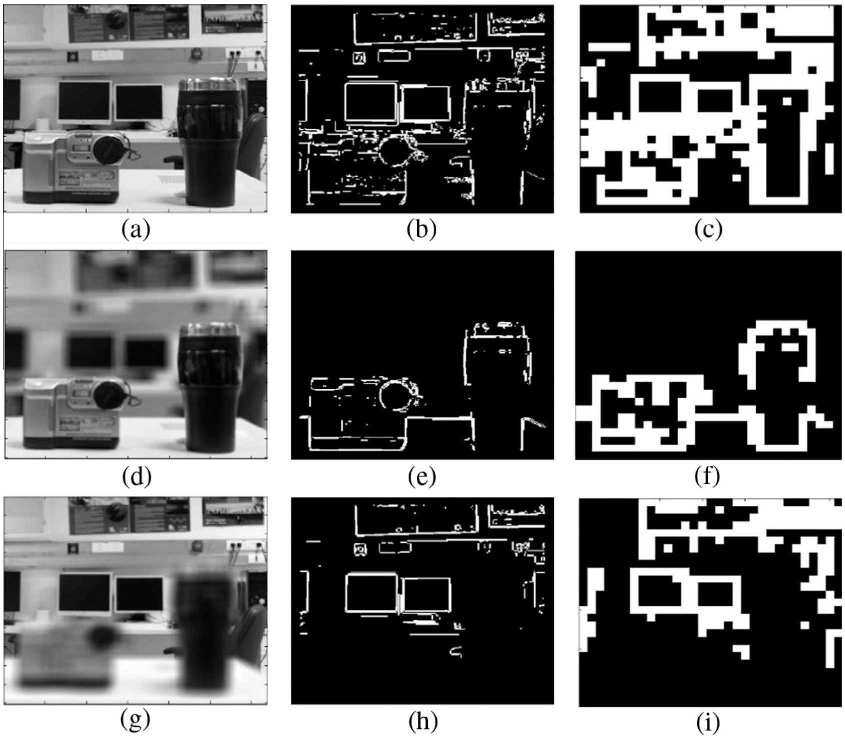


Fig. 16. Results of automatic OI depth plane selection with confocal de-cluttering using a light-field setup. Top row (a–c) shows the center view image, together with the edge image and its resolution-compressed version. Middle row (d–f) shows the same images for the confocal image reconstructed at the 0.6 m distance identified by the detection algorithm. The bottom row (g–i) shows the same results for reconstruction at the other local peak distance of 3 m.

cussion, with the use of scanning, observers can learn to elicit specific response patterns in multiple choice situations that they can discriminate (Bionic Vision Australia, 2014; Second Sight Europe, 2013). However, such performance is not likely to be generalized to the recognition of patterns not previously learned.

Our results may appear to suggest that the resolution needed for object recognition at the 50% correct level, even with our proposed confocal imaging, is much higher than what is achievable with current prostheses, and even higher than the anticipated resolution of the next few generations of such systems. It seems that a resolution of 3,000–5,000 electrodes may be needed. This estimate may be overly pessimistic, as we used static images in our testing. Most current and anticipated visual prosthetic devices use video motion imaging (although the frame rate is usually at 10 frames per second or lower). With live video, not only motion imaging is provided but performance can be much improved. The small variation in the input due to electronic noise and to slight jitter in camera position due to head tremor and bobbing result in slightly different images being acquired and processed at each frame, even when examining a static object while sitting (Peli & Garcia-Perez, 2003). At the low resolutions and dynamic ranges we deal with here, that effect may result in temporal averaging of the signal that filters away some of the noise and pulls out the consistent visual signal. This is an effect similar to stochastic resonance (Collins, Imhoff, & Grigg, 1996). Improvement in resolution with image jitter was recently demonstrated for patients with AMD (Watson et al., 2012) and was simulated for bipolar (3 level) visual edge detection (Peli, 2002). Bipolar edges can be implemented if the dynamic range of visual prostheses is improved beyond the current 1 bit level. Super-resolution benefits were also suggested using dynamic halftones (Mulligan, Ahumada, & Jr., 1992). Many years ago, when computer displays had only 2 bits of gray scale, we demonstrated that quartertone coding can provide a substantial resolution benefit over binary imaging (Goldstein, Peli, & Wooledge, 1987). We noted during our object recognition trials that when subjects failed to recognize objects they frequently rotated and shifted their head as if trying to generate different viewpoints or motion parallax, intuitively attempting to separate object from background. This was unhelpful in our experimental setting but would likely improve performance if applied in a motion video system. Thus, with video imaging the performance could be improved, and may reach an acceptable level at a lower, more practical, prosthetic resolution.

It is interesting to consider our object recognition task results in the context of threshold performance. The lateral separation between the psychometric functions fitting the individual subjects' data for both conditions (background cluttered and de-cluttered) (Fig. S2 in supplement), as well as the cumulative performance of all of the subjects (Fig. 8), diminishes as the threshold is increased. This may be considered a limitation of our approach, as one cares more about the impact of the confocal imaging at higher levels of performance, which are more desirable than in lower levels of performance. However, it is important to realize that even performance in an object recognition task at a level of 50% correct would be highly desirable for any current visual prosthesis. If and when the performance level reaches as high as 90% correct, diminishing effects of our proposed confocal imaging may not be too much of a loss.

Although our preliminary experiment was sufficient to show the significance of the background clutter effect on object recognition at low resolution, further aspects should be considered in future work. Here we used simple objects, yet recognition varied between subjects, as some objects were difficult to recognize from just the basic shape. For example, the cylindrical shape of the desk lamp (object 11) was easily recognized by subjects, but they could only identify the partial shape and not the whole item. In addition, linguistic analysis of subjects' responses was required to decide correct answers. A more systematic methodology is required to create a dataset of

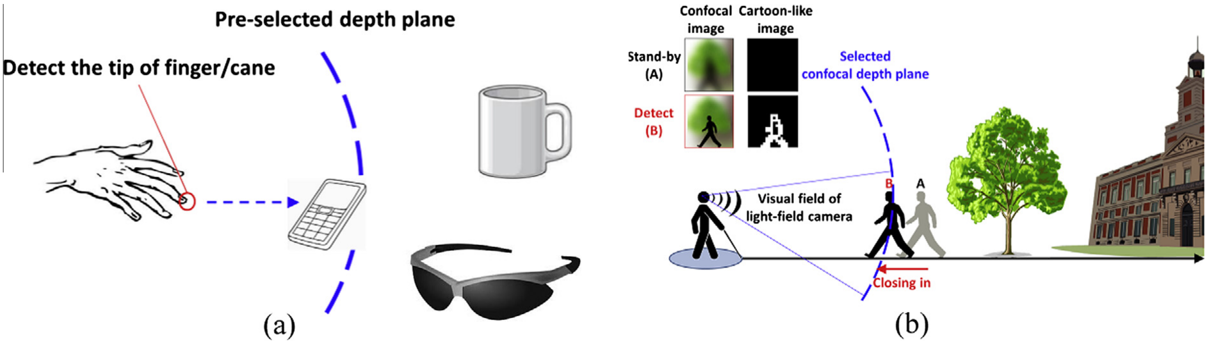


Fig. 17. Operating modes of active confocal imaging for visual prostheses (a) Confocal-extension mode. The user, trying to find an object, reaches and touches around the area where the object is expected to be. The system first detects the tip of the finger or cane and sets the focal distance to a predefined distance in front of it. In this mode, users can see farther than the arm or cane length, hence the designation confocal-extension, and we expect this extended search range to reduce the search time. (b) Obstacle avoidance mode, to be used mainly when walking. The system displays only objects that enter the pre-selected distance range and will alert the user when such an object is detected (moving from location A to B in the figure). The range included may be selected to be very narrow or wider. This mode calls attention to obstacles or hazards that are missed or not reachable by the cane. When an obstacle is detected the user may execute an avoidance maneuver based on the "visual" information displayed.

images that can be used even more reliably in such studies. The use of light-field imaging may support the use of such a dataset by numerous groups testing different prostheses, as it contains the full 3D information and thus may be used in video and in conjunction with the head movement and other depth cues.

As shown in Figs. 6 and 7, and in Fig 16e and f, the diagonal edge from the ground plane (desk edge, in this case) clutters the OIs in both background cluttered and de-cluttered conditions. Even if the confocal image captures only narrow depth and the focal distance is set on the OI accurately, the ground plane around the focal distance is also captured in focus and it is not removed by the current de-cluttering method. Frequently, a shadow of the object projected on the ground plane may have enough contrast and sharpness to be maintained. A de-cluttering process that detects and removes the ground plane may further improve the performance of the system.

Light-field cameras are already on the market from Lytro and Raytrix. Pelican Imaging (Mountain View, CA) (Venkataraman, Jabbi, & Mullis, 2011) and Toshiba (Tokyo, Japan) (Kamiguri, 2012) are developing modules for smart phones. These modules will be easily adaptable for visual prosthetic vision use and are expected to be inexpensive. In future work, a light-field camera will have to be used in video mode. An inexpensive commercial light-field camera (Lytro) exists but is not suitable for visual prosthesis applications. The field of view in Lytro can be adjusted from 5° to 40°. However, the confocal performance (DOF) at the wide field of view setting is too broad to suppress background clutter (Min, Kim, & Lee, 2005). The optimal field of view of the Lytro camera is 8°, where it can generate a maximum of 7 confocal images at depth planes of 10, 25, 50, 100, 200, 600, and 1100 cm (Maiello, 2013), sufficient DOF ranges for our application but insufficient visual field. At wider field of view settings the confocal depth steps are too sparse to suppress clutter in other planes. The other commercial light-field camera (Raytrix) can be customized by optimizing the lens array design for confocal imaging of the wide-field light-field. It can operate at video rate and have a narrower DOF, but it is much more expensive with the customization option. Yet, it can support evaluation of the feasibility and the utility of such a system.

An active confocal imaging system offers many possibilities for modes of operation in future prostheses. Obvious candidates include a free-search mode (Animation 4 in the online supplement), which would be especially useful for orientation. A controller mounted on the handle of a long cane could be used to isolate and then zoom in on one of several objects selected automatically from the image obtained by the head-mounted camera. Another mode, confocal-extension mode (Fig. 17a), may be useful for finding objects slightly beyond the reach of the arm or cane. The confocal depth would be set to a narrow band, and the presented de-cluttered view would be centered laterally based on the detected location of the user's searching hand or the tip of the long cane. Further, in an obstacle-avoidance mode (Fig. 17b), objects in the oncoming path could be detected when reaching a preset distance from the user and presented visually, together with an audible or haptic alert, giving sufficient warning and providing for visually guided avoidance maneuvers. This mode may be especially useful for elevated obstacles that the long cane cannot detect, and as an early warning of pedestrian traffic (Pundlik, Tomasi, & Luo, 2014). Active vision, where the user selects the mode of operation and interacts with the environment through the prosthesis settings, is our preference, in contrast to other computer-vision based approaches. Our approach counts on the user's knowledge of the environment and awareness of what he wishes to achieve, rather than on a need for the system to be able to guess it.

Acknowledgments

This work was supported in part by NIH Grant R01EY05957. Jung and Peli have applied for a patent on the active confocal imaging system for visual prostheses. If a patent is awarded, it will be assigned to the Schepens Eye Research Institute.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2014.10.023.

References

Ahuja, A. K., & Behrend, M. R. (2013). The Argus II retinal prosthesis: Factors affecting patient selection for implantation. *Progress in Retinal and Eye Research, 2*(4), 1–15.

Al-Atabany, W., McGovern, B., Mehran, K., Berlinguer-Palmini, R., & Degenaar, P. (2013). A processing platform for optoelectronic/optogenetic retinal prosthesis. *IEEE Transactions on Biomedical Engineering, 60*(3), 781–791.

Aloni, D., & Yitzhaky, Y. (2014). Detection of object existence from a single reconstructed plane obtained by integral imaging. *IEEE Photonics Technology Letters, 26*(7), 726–728.

American Foundation for the Blind (2011*). Statistical snapshots from the American Foundation for the blind.* <http://www.afb.org/Section.asp?SectionID=15> Accessed 16.06.

Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology, 51*(10), 809–816.

Bionic Vision Australia (2014). *Dianne Ashworth bionic eye prototype testing, 2014*. <https://www.youtube.com/watch?v=6EmleCs0KGY> Accessed.

Boyer, K. L., & Kak, A. C. (1987). Color-encoded structured light for rapid active ranging. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9*(1), 14–28.

Brown, M. M., Brown, G. C., Sharma, S., Kistler, J., & Brown, H. (2001). Utility values associated with blindness in an adult population. *British Journal of Ophthalmology, 85*(3), 327–331.

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

Chen, S. C., Suaning, G. J., Morley, J. W., & Lovell, N. H. (2009b). Simulating prosthetic vision: II. Measuring functional capacity. *Vision Research, 49*(19), 2329–2349.

Chéné, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., Morel, P., et al. (2012). On the use of depth camera for 3D phenotyping of entire plants. *Computers and Electronics in Agriculture, 82*, 122–127.

Chouvardas, V. G., Miliou, A. N., & Hatalis, M. K. (2008). Tactile displays: Overview and recent advances. *Displays, 29*(3), 185–194.

Collins, J. J., Imhoff, T. T., & Grigg, P. (1996). Noise-enhanced tactile sensation. *Nature, 383*(6603), 770.

Cooke, J. (2006). *Tongue technology aids blind*. <http://news.bbc.co.uk/player/nol/newsid_6170000/newsid_6170500/6170531.stm?bw=bb&mp=wm> Accessed.

da Cruz, L., Coley, B. F., Dorn, J., Merlini, F., Filley, E., Christopher, P., et al. (2013). The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss. *British Journal of Ophthalmology, 97*(5), 632–636.

Dowling, J. A., Maeder, A., & Boles, W. (2004). Mobility enhancement and assessment for a visual prosthesis. *Proceedings of the SPIE, 5369*, 780–791.

El-laithy, R. A., Jidong, H., & Yeh, M. (2012). Study on the use of Microsoft Kinect for robotics applications. *Proceedings of the IEEE Position Location and Navigation Symposium (PLANS)*, 1280–1288.

Goldstein, R. B., Peli, E., & Wooledge, K. (1987). Medical image communication using halftone algorithms. *Proceedings of the Society of Photo-Optical Instrumentation Engineers, 845*, 413–418 (Boston, MA).

Hao, T., Ro, T., & Zhigang, Z. (2013). Smart sampling and transducing 3D scenes for the visually impaired. *Proceedings of the Multimedia and Expo Workshops (ICMEW)*, 1–4.

Harris, M. (2012). Light-field photography revolutionizes imaging. *IEEE Spectrum, 49*(5), 44–50.

Hong, S. H., Jang, J. S., & Javidi, B. (2004). Three-dimensional volumetric object reconstruction using computational integral imaging. *Optics Express, 12*(3), 483–491.

Horowitz, A. (2004). The prevalence and consequences of vision impairment in later life. *Topics in Geriatric Rehabilitation, 20*(3), 185–195.

Horsager, A., Greenberg, R. J., & Fine, I. (2010). Spatiotemporal interactions in retinal prosthesis subjects. *Investigative Ophthalmology & Visual Science, 51*(2), 1223–1233.

Humayun, M. S., Dorn, J. D., da Cruz, L., Dagnelie, G., Sahel, J.-A., Stanga, P. E., et al. (2012). Interim results from the international trial of second sight's visual prosthesis. *Ophthalmology, 119*(4), 779–788.

Hwang, A. D., & Peli, E. (2014). An augmented-reality edge enhancement application for Google Glass. *Optometry and Vision Science, 91*(8), 1021–1030.

Jung, J. H., Hong, K., Park, G., Chung, I., Park, J. H., & Lee, B. (2010). Reconstruction of three-dimensional occluded object using optical flow and triangular mesh reconstruction in integral imaging. *Optics Express, 18*(25), 26373–26387.

Jung, J.-H., Park, S.-G., Kim, Y., & Lee, B. (2012). Integral imaging using a color filter pinhole array on a display panel. *Optics Express, 20*(17), 18744–18756.

Kamiguri, T. (2012). *Toshiba putting focus on taking misfocusing out of photos*. <http://ajw.asahi.com/article/business/AJ201212270054> Accessed.

Kim, D.-H., Erdenebat, M.-U., Kwon, K.-C., Jeong, J.-S., Lee, J.-W., Kim, K.-A., et al. (2013). Real-time 3D display system based on computer-generated integral imaging technique using enhanced ISPP for hexagonal lens array. *Applied Optics, 52*(34), 8411–8418.

Kim, Y., Hong, K., & Lee, B. (2010). Recent researches based on integral imaging display method. *3D Research, 1*(1), 17–27.

Kim, J., Jung, J.-H., Jang, C., & Lee, B. (2013). Real-time capturing and 3D visualization method based on integral imaging. *Optics Express, 21*(16), 18742–18753.

Kim, J., Jung, J.-H., Jeong, Y., Hong, K., & Lee, B. (2014). Real-time integral imaging system for light field microscopy. *Optics Express, 22*(9), 10210–10220.

Kuyk, T., Liu, L., Elliott, J., Grubbs, H., Owsley, C., McGwin, G. Jr.,, et al. (2008). Health-related quality of life following blind rehabilitation. *Quality of Life Research, 17*(4), 497–507.

Lange, R., & Seitz, P. (2001). Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics, 37*(3), 390–397.

Lee, K.-J., Hwang, D.-C., Kim, S.-C., & Kim, E.-S. (2008). Blur-metric-based resolution enhancement of computationally reconstructed integral images. *Applied Optics, 47*(15), 2859–2869.

Levoy, M., Chen, B., Vaish, V., Horowitz, M., McDowall, I., & Bolas, M. (2004). *Synthetic aperture confocal imaging, ACM SIGGRAPH 2004 papers*. Los Angeles, California: ACM, pp. 825–834.

Li, W. H. (2013). Wearable computer vision systems for a cortical visual prosthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) workshops*.

Lieby, P., Barnes, N., McCarthy, C., Nianjun, L., Dennett, H., Walker, J. G., Botea, V., & Scott, A. F. (2011). Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions (pp. 8017–8020).

Lippmann, G. (1908). La photographie integrale. *Comptes Rendus de l'Académie des Sciences, 146*, 446–451.

Maiello, G. (2013). *The contribution of defocus blur to depth perception through stereopsis*. Unpublished Master's, University of Genoa Polytechnic School.

Mallat, S. G. (1989). Multifrequency channel decompositions of images and the wavelet models. *IEEE, 37*(12), 2091–2110.

Margalit, E., Maia, M., Weiland, J. D., Greenberg, R. J., Fujii, G. Y., Torres, G., et al. (2002). Retinal prosthesis for the blind. *Survey of Ophthalmology, 47*(4), 335–356.

McCarthy, C., Barnes, N., & Lieby, P. (2011). Ground surface segmentation for navigation with a low resolution visual prosthesis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011*, 4457–4460.

Mcnamara, M. (2007). *Technology may give blind a touch of sight*. <http://www.cbsnews.com/news/technology-may-give-blind-a-touch-of-sight/> Accessed.

Min, S.-W., Jung, S., Park, J.-H., & Lee, B. (2001). Three-dimensional display system based on computer-generated integral photography. *Proceedings of the SPIE, 4297*, 187–195.

Min, S.-W., Kim, J., & Lee, B. (2005). New characteristic equation of three-dimensional integral imaging system and its applications. *Japanese Journal of Applied Physics, 44*(2), L71–L74.

Mulligan, J. B., & Ahumada, A. J., Jr. (1992). Principled halftoning based on human vision models. In *Proceedings of human vision, visual processing and digital display III*, 1666, Bellingham, WA (pp. 109–121).

Nau, A., Bach, M., & Fisher, C. (2013). Clinical tests of ultra-low vision used to evaluate rudimentary visual perceptions enabled by the BrainPort Vision Device. *Translational Vision Science & Technology, 2*(3), 1.

Neve, C. D. (2011). *Jose Neto: Search for sight*. <https://www.youtube.com/watch?v=_nBs7PnKxzE> Accessed.

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., & Hanrahan, P. (2005). *Light field photography with a hand-held plenoptic camera*. Stanford University.

Ong, J. M., & Cruz, L. d. (2012). The bionic eye: A review. *Clinical & Experimental Ophthalmology, 40*(1), 6–17.

Palanker, D., Vankov, A., Huie, P., & Baccus, S. (2005). Design of a high-resolution optoelectronic retinal prosthesis. *Journal of Neural Engineering, 2*(1), S105–S120.

Parikh, N. J., McIntosh, B. P., Tanguay, A. R., Jr., Humayun, M. S., & Weiland, J. D. (2009). Biomimetic image processing for retinal prostheses: Peripheral saliency cues. In *Proceedings of the 31st annual international conference of the IEEE engineering in medicine and biology society*, Minneapolis, MN (pp. 4569–4572).

Parikh, N., Itti, L., & Weiland, J. (2010). Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering, 7*(1), 16006.

Park, J.-H., Hong, K., & Lee, B. (2009). Recent progress in three-dimensional information processing based on integral imaging. *Applied Optics, 48*(34), H77–H94.

Peli, E. (2002). Feature detection algorithm based on a visual system model. *Proceedings of the IEEE, 90*(1), 78–93.

Peli, E., & Garcia-Perez, M. A. (2003). Motion perception during involuntary eye vibration. *Experimental Brain Research, 149*(4), 431–438.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364), 699–705.

Pundlik, S., Tomasi, M., & Luo, G. (2014). Evaluation of a portable collision warning device for visually impaired patients in an indoor obstacle course (abstract). *Investigative Ophthalmology & Visual Science*. E-abstract 2154.

Rizzo, J. F., 3rd, Wyatt, J., Loewenstein, J., Kelly, S., & Shire, D. (2003a). Methods and perceptual thresholds for short-term electrical stimulation of human retina with microelectrode arrays. *Investigative Ophthalmology & Visual Science, 44*(12), 5355–5361.

Rizzo, J. F., 3rd, Wyatt, J., Loewenstein, J., Kelly, S., & Shire, D. (2003b). Perceptual efficacy of electrical stimulation of human retina with a microelectrode array during short-term surgical trials. *Investigative Ophthalmology & Visual Science, 44*(12), 5362–5369.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*(2).

Second Sight Europe (2013). *The Argus II retinal implant allows letter, word reading in patients with blindess*. <https://www.youtube.com/watch?v=YU1F4TcGRlQ> Accessed.

Second Sight Medical Products Inc. (2013). *Argus II retinal prosthesis system surgeon manual* (Vol. 090001-004).

Singer, M. A., Amir, N., Herro, A., Porbandarwalla, S. S., & Pollard, J. (2012). Improving quality of life in patients with end-stage age-related macular degeneration: Focus on miniature ocular implants. *Clinical Ophthalmology, 6*, 33–39.

Sobel, I., & Feldman, G. (1968). A 3x3 isotropic gradient operator for image processing. Presented at the *Stanford Artificial Intelligence Project (SAIL)*.

Stern, A., & Javidi, B. (2006). Three dimensional sensing, visualization, and processing using integral imaging. *Proceedings of IEEE, Special Issue on 3D Technologies for Imaging and Display, 94*, 591–607.

Stingl, K., Bartz-Schmidt, K. U., Gekeler, F., Kusnyerik, A., Sachs, H., & Zrenner, E. (2013). Functional outcome in subretinal electronic implants depends on foveal eccentricity. *Investigative Ophthalmology & Visual Science, 54*(12), 7658–7665.

Venkataraman, K., Jabbi, A. S., & Mullis, R. H. (2011). Capturing and processing of images using monolithic camera array with heterogeneous imagers, US20110069189 A1.

Wang, L., Yang, L., & Dagnelie, G. (2008). Virtual wayfinding using simulated prosthetic vision in gaze-locked viewing. *Optometry and Vision Science, 85*(11), 1057–1063.

Ward, J., & Meijer, P. (2010). Visual experiences in the blind induced by an auditory sensory substitution device. *Consciousness and Cognition, 19*(1), 492–500.

Watson, L. M., Strang, N. C., Scobie, F., Love, G. D., Seidel, D., & Manahilov, V. (2012). Image jitter enhances visual performance when spatial resolution is impaired. *Investigative Ophthalmological & Visual Science, 53*(10), 6004–6010.

Weiland, J. D., Cho, A. K., & Humayun, M. S. (2011). Retinal prostheses: Current clinical results and future needs. *Ophthalmology, 118*(11), 2227–2237.

Weiland, J. D., Parikh, N., Pradeep, V., & Medioni, G. (2012). Smart image processing system for retinal prosthesis. In *Conference proceedings of the 34th annual international conference of the IEEE engineering in medicine and biology society, 2012* (pp. 300–303).

Wichmann, F., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Wilke, R. H., Moghaddam, G., Dokos, S., Suaning, G., & Lovell, N. (2010). Stimulation of the retinal network in bionic vision devices: From multi-electrode arrays to pixelated vision. In K. Wong, B. S. Mendis, & A. Bouzerdoum (Eds.). *Neural information processing. Theory and algorithms* (Vol. 6443, pp. 140–147). Berlin Heidelberg: Springer.

World Health Organization (2013). *Visual impairment and blindness*. <http://www.who.int/mediacentre/factsheets/fs282/en/> Accessed.

Yitzhaky, Y., & Peli, E. (2003). A method for objective edge detection evaluation and detector parameter selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(8), 1027–1033.

Zhao, Y., Lu, Y., Tian, Y., Li, L., Ren, Q., & Chai, X. (2010). Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision. *Information Sciences, 180*(16), 2915–2924.

Zrenner, E., Bartz-Schmidt, K. U., Benav, H., Besch, D., Bruckmann, A., & Gabel, V.-P. (2011). Subretinal electronic chips allow blind patients to read letters and combine them to words. *Proceedings of the Royal Society B: Biological Sciences, 278*(1711), 1489–1497.

Supplementary Material for: Active Confocal Imaging for Visual Prostheses

Figures

1

2

3

4

5

6

7

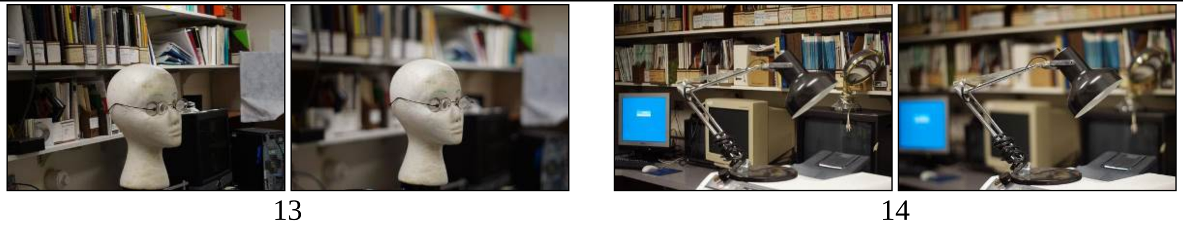
8

9

10

11

12



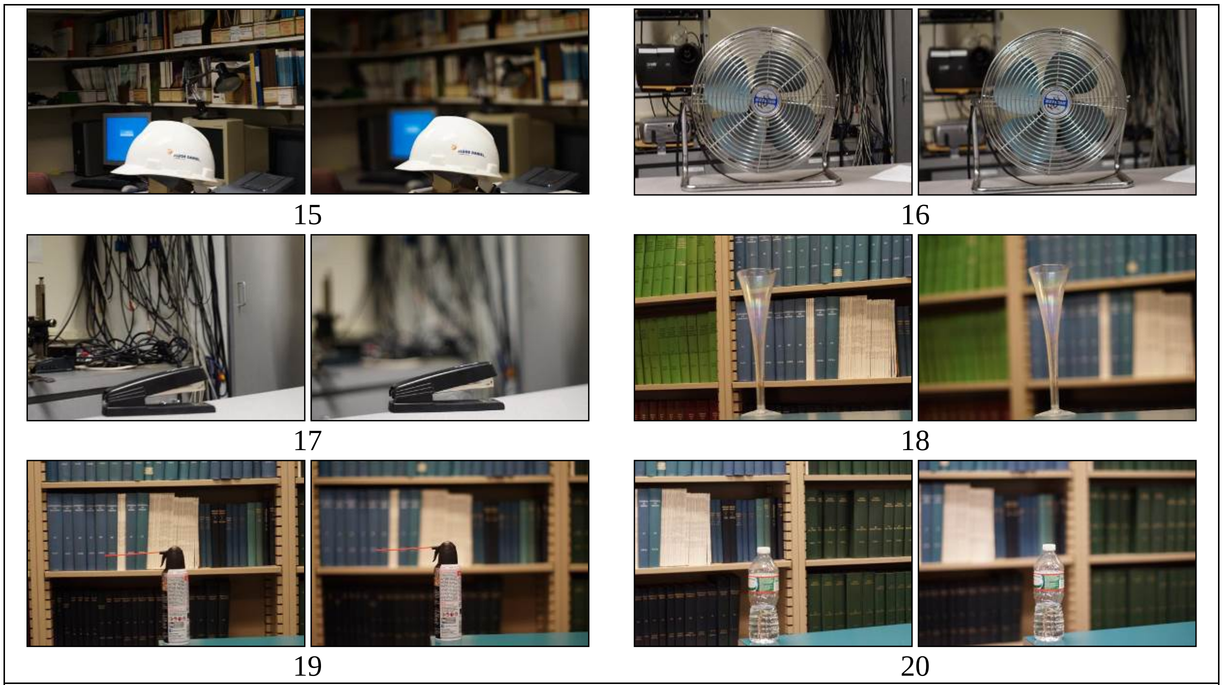


Figure S1. Dataset of objects for subject test (conventional wide DOF on the left and confocal narrow DOF images on the right). Numbers indicate the object number mentioned in the paper.

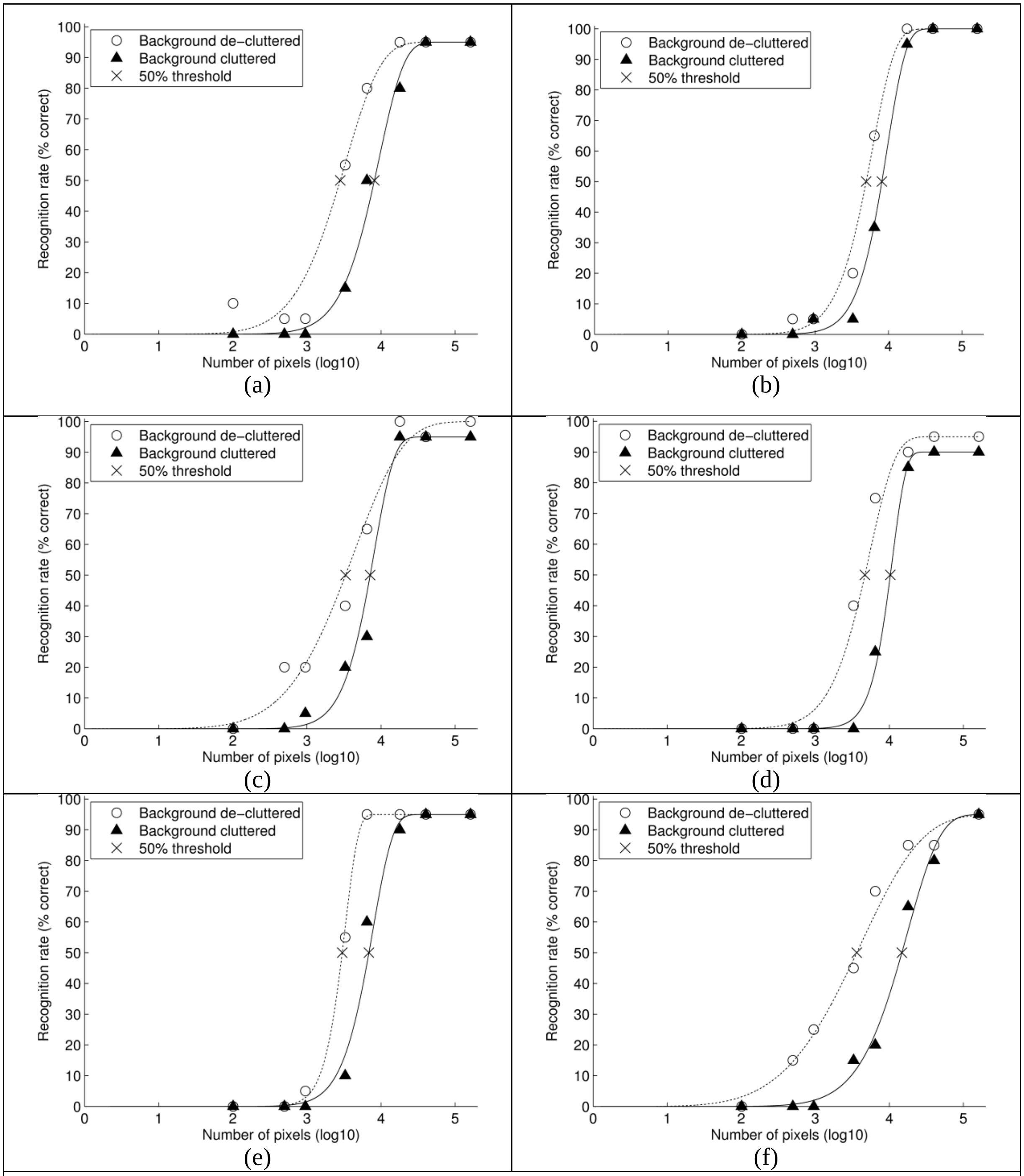


Figure S2. Object recognition rate for individual subjects. Data for each resolution level are fitted with a Weibull function for each condition.

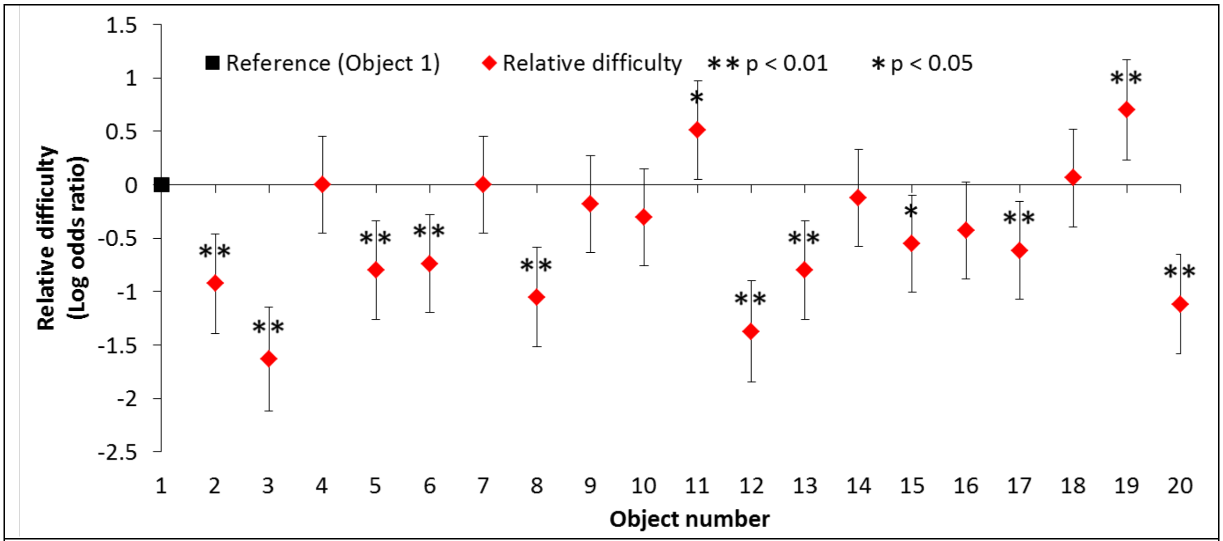


Figure S3. Relative difficulty of recognizing objects in the dataset compared with object 1. With the recognition difficulty of object 1 set as 0 for reference, relative difficulties of other objects compared with object 1 are analyzed by the binary logistic regression model and represented here by log odds ratio. Error bars are 95% confidence intervals derived from the standard errors of the logistic regression coefficients. Significance marks * and ** show α < 0.05 and α < 0.01, respectively. Because the reference object, selected arbitrarily, is relatively difficult to recognize, difficulties of most objects are negative (easier).

Figure S3 shows the relative difficulty of each object in the dataset compared with the reference object (object 1). We use log odds ratios instead of odds ratios to intuitively identify easier and more difficult objects to be recognized. For example, the relative difficulty of the object 6 is 0.18, which means that object 6 is about 5 times easier than the reference object. The relative difficulty of object 19 is 5.0, opposite the case for object 6, which means that it is 5 times more difficult to be recognized. However, the difference cannot be shown intuitively if we plot it without the conversion to log units.

Because object 1 was a relatively difficult object to recognize, with unclosed edge lines and clutter of letters, the difficulties of objects in dataset were mostly lower than the reference. For example, object 3 is the easiest object to recognize, with a clear outline, as shown in the compressed image in Fig. 5. However, objects 11 and 19 were difficult for subjects to recognize. In the case of object 11, subjects could not identify it as a lamp because of its unusual shape, although they could easily describe its cylindrical shape. For object 19, the nozzle of the spray can was mainly ignored by subjects, and they usually identified it as a bottle or a can because they thought the nozzle was part of the background and noise in the edge detection. Although this analysis is not sufficient with the small sample to categorize the dataset, and generating a final data set was not the purpose of this experiment, at least this result shows that difficulties are moderately balanced.