Information Sciences 277 (2014) 512–524

Contents lists available at ScienceDirect

IranPaper.blog.ir

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

Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies

Jing Wang [a], Yanyu Lu [b], Liujun Gu [a], Chuanqing Zhou [a], Xinyu Chai [a,*]

[a] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 2 August 2012
Received in revised form 14 January 2014
Accepted 25 February 2014
Available online 4 March 2014

Keywords:
Visual prosthesis
Simulated prosthetic vision
Background subtraction
Image processing strategy
Moving objects recognition

ABSTRACT

A visual prosthesis that applies electrical stimulation to different parts of the visual pathway has been proposed as a viable approach to restore functional vision. However, the created percept is currently limited due to the low-resolution images elicited from a limited number of stimulating electrodes. Thus, methods to optimize the visual percepts providing useful visual information are being considered. We used two image-processing strategies based on a novel background subtraction technique to optimize the content of dynamic scenes of daily life. Psychophysical results showed that background reduction, or background reduction with foreground enhancement, increased response accuracy compared with methods that directly merged pixels to lower resolution. By adding more gray scale information, a background reduction/foreground enhancement strategy resulted in the best performance and highest recognition accuracy. Further development of image-processing modules for a visual prosthesis based on these results will assist implant recipients to avoid dangerous situations and attain independent mobility in daily life.

1. Introduction

There are no effective clinical treatments to restore vision for some retinal diseases such as age-related macular degeneration and retinitis pigmentosa. Implanting a visual prosthesis has been proposed as a viable approach to restore partial vision to blind patients suffering from these diseases. The perception of spots of light, called phosphenes, are elicited by electrically stimulating different parts of the visual pathway (retina, optic nerve, or cortex) [42,47]. Over recent decades, several research groups have developed different types of visual prosthetic devices and successfully implanted them in blind patients [7,27,40,48]. Although visual prostheses have gained significant development and continue to achieve encouraging improvement, some engineering challenges, such as electrode fabrication, power consumption, and long-term viability [23,42], remain to be overcome before microelectronic high density electrode implants can be realized. Consequently, visual perception generated by a limited number of stimulation contacts is still poor relative to normal vision. Methods to optimize the image quality presented by such a limited number of phosphene dots to maximize visual percepts are currently being considered.

* Corresponding author. Tel.: +86 021 34204077; fax: +86 021 34204078.
E-mail address: xychai@sjtu.edu.cn (X. Chai).

http://dx.doi.org/10.1016/j.ins.2014.02.136
0020-0255/

Simulation of prosthetic vision offers an alternative way of adjusting implant designs and estimate the minimal information requirements for the prosthetic wearer. Currently used models of prosthetic vision simulate phosphene shape, intensity, size and regularity to simplify the possible range of percepts by normally sighted volunteers, and therefore, represent a best case scenario for a blind patient under these conditions [15]. In this way, many research groups have effectively estimated the performance capacity of a visual prosthesis for such things as facial recognition [36] and eccentric reading [33]. In order to provide prosthesis wearers with useful artificial vision, some researchers have incorporated and estimated several image processing strategies applicable to some basic visual functions in addition to the essential processing of image downsampling and pixelization. Boyle et al. [5] studied the effect of image processing factors on object and face recognition, which included spatial resolution, grayscale, contrast, edge detection, distance and importance mapping, in order to enhance the information content presented by limited numbers of electrodes. Dowling et al. [22] reviewed aspects related to blind mobility and discussed a variety of image processing techniques suitable for a visual prosthesis, and presented a mobility display framework. Based on previous work, Boyle et al. [6] applied several variations to region-of-interest (ROI) processing and edge detection for scene and face recognition. Results showed that presenting a digital zoom of a salient region within the generated ROI was preferable to presenting an entire ROI-processed image. Zhao et al. [44] used two kinds of image processing strategies (adaptive threshold-binarization and edge extraction) to investigate the effect of pixel shape and resolution in a recognition task of common objects or scenes. Results demonstrated that image mode had a significant impact on recognition accuracy near the threshold of recognition. van Rheede and colleagues [39] have implemented three techniques (Full-field representation, ROI and Fisheye) for optimizing the information content of a simulated prosthetic image in order to evaluate visual acuity, object recognition and manipulation, and navigation within the environment. Their experimental results suggested that changing the conditions of image presentation proved advantageous for different visual functional tasks.

Static object and scene recognition have been extensively studied, however, the capacity to perceive moving objects under circumstances such as danger avoidance and independent mobility, which is an important aspect of vision [17], have rarely been investigated with simulated prosthetic vision. McCarthy and Barnes [32] proposed a time-to-contact map based on depth image by focusing on free-moving incoming object perception. Their quantitative (sequence analysis) and qualitative (image presentations) simulations demonstrated the effectiveness of the proposed representation method for emphasizing objects posing an imminent threat of collision (via increased phosphene brightness). To facilitate the perception of moving objects in a general dynamic scene by a prosthesis wearer, we used two image processing strategies based on a universal background subtraction technique with an adaptive post-processing method. The effect and processing speed of these strategies were evaluated by a series of psychophysical experiments on the perception of a dynamic scene. We demonstrate the usefulness of background-subtraction-based image processing strategies for recognition of moving objects that simulate the experience of a blind patient implanted with a prosthesis.

2. Material and methods

2.1. Materials

Ten different dynamic scenarios (five indoors, five outdoors) were filmed and included images of people during daily activities or different kinds of moving vehicles, i.e. situations that visually impaired people might expect to routinely encounter. Each video sequence was ~10 s in duration, in which the vertical viewing angle of every object (human or vehicle) was ~ equal. A 20° × 20° field of view was adopted based on the available visual field of a retinal prosthesis prototype [46]. Indoor scenes showed people presented in a vertical view that covered a visual angle of 16°–18°, whereas moving objects in an outdoor scene were presented within a 3°–5° visual angle so that all the observed moving objects remained within the entire 20° field of view. The playing speed of all scenes was set to 20 frames per second (examples of screenshots from the original video materials are shown in Fig. 1).

2.2. Image processing strategies

RGB color video images were resized to a 480 × 480 resolution, converted to grayscale, and then adjusted for uniform illumination. In general, the image processing stage of the visual prosthesis adjusts the image resolution by combining a set number of pixels into a single output pixel for stimulating the tissue interface array [18] and is called Merging Pixels to Low Resolution (MPLR). The small electrode number leads to a huge loss of information when presenting a real-life scene. Increasing the contrast between a moving object(s) (i.e. the foreground) and the relatively static or slow moving parts of the scene (i.e. the background) can enhance the perception of the main information. Therefore, moving objects in a dynamic scene need to be automatically detected and precisely separated from the surrounding information, i.e. a series of foreground segmentation images must be generated first from the original video. A common motion detection technique for distinguishing the foreground from the background in computer vision is background subtraction (BS). We applied a novel BS algorithm for foreground extraction and developed a post-processing method for optimizing the segmentation. Two image-processing strategies based on BS segmentation were used to increase the contrast between the moving foreground and background (Fig. 2), and then compared with direct MPLR processing.

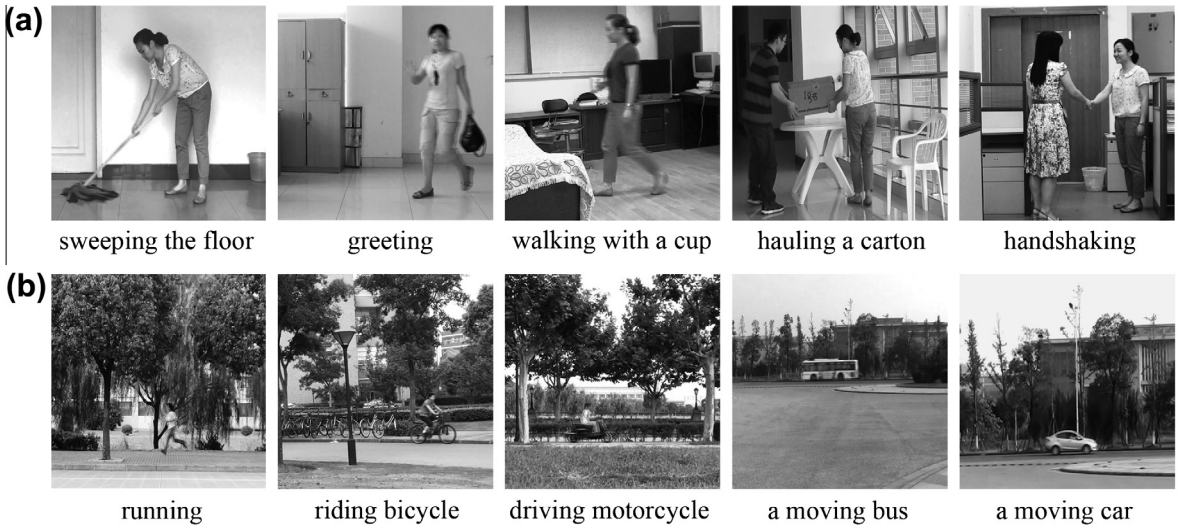

Fig. 1. Screenshots taken from each of the original video sequences. (a) Example frames from indoor videos containing images of daily activity. (b) Examples of outdoor videos containing continuous movement by different vehicles and/or people.

2.2.1. Background subtraction (BS)

2.2.1.1. Visual background extractor (ViBe). BS regards the foreground as the difference between the current image and a reference background image, often called the "background model". Several techniques have been used, such as the single Gaussian model [43] or temporal median filtering [1,31] that are simple models but extremely sensitive to changes in dynamic scenes resulting from lighting and/or extraneous events. The Gaussians Mixture Model [34] and its enhanced version [29] in real-time tracking provide good model accuracy, albeit at the price of increased algorithm complexity and computational time. A recent proposed universal BS technique called visual background extractor (ViBe) outperforms other techniques, not only in terms of computational speed and classification accuracy [2,3], but also the ability to solve problems arising from the nature of video surveillance [8]. Due to its excellent performance and almost parameterless procedure, we used ViBe for foreground segmentation. ViBe compares current input pixels with a background model by randomly selecting neighboring pixels.

The ViBe background model is initialized with the neighborhood of each observed pixel from the first frame based on the assumption that a pixel and its neighbors share a similar distribution in the spatial domain. Suppose at the starting time in a video sequence that $BG^1(x,y)$, $N_G^1(x,y)$, $(x,y)_m^1$ respectively represent the initial background model, the spatial neighborhood of the pixel, and pixel samples in the model from the first frame. The initial background model is defined as:

$$BG^1(x,y) = \left\{ (x,y)_m^1 \middle| (x,y)_m^1 \in N_G^1(x,y), \quad m = 1,2,\ldots,N \right\} \tag{1}$$

$(x,y)_m^1$ is randomly selected in $N_G^1(x,y)$ according to uniform probability, and N is the total number of samples.

When $t = k$ ($k \neq 1$), each $(x,y)^k$ in frame $f(x,y)^k$ is estimated via the last established background model $BG^{k-1}(x,y)$ at this location. In order to judge whether the pixel $(x,y)^k$ belongs to the background, the following equations are used:

$$S_R^k = \left\{ (x,y)_R^{k-1} \middle| \|(x,y)^k - (x,y)_R^{k-1}\|_{Euclid} \leqslant R, (x,y)_R^{k-1} \in BG^{k-1}(x,y) \right\} \tag{2}$$

$$\#\{S_R^k\} \leqslant \#_{\min} \tag{3}$$

where R is a spherical searching scope of $(x,y)^k$. S_R^k is the pixel set in which the distance between each pixel $(x,y)_R^{k-1}$ in $BG^{k-1}(x,y)$ and $(x,y)^k$ is less than R. If the cardinality of S_R^k, i.e. $\#\{S_R^k\}$, is less than a threshold $\#_{\min}$, $(x,y)^k$ is classified as the background pixel. Otherwise, the pixel belongs to foreground. Classification processing ends when $\#_{\min}$ matches are found.

Due to the spatiotemporal property of background subtraction, a memoryless update policy was adopted for discarding some pixel samples in the model. The probability P of preserving a sample present in the model at time t after the update of the pixel model is given by $(N-1)/N$, which is denoted as

$$P(t,t+dt) = \left(\frac{N-1}{N}\right)^{(t+dt-t)} \tag{4}$$

We first took the logarithm of both sides in the equation and then exponentially transformed it. After rearranging the terms, Eq. (4) can be expressed as

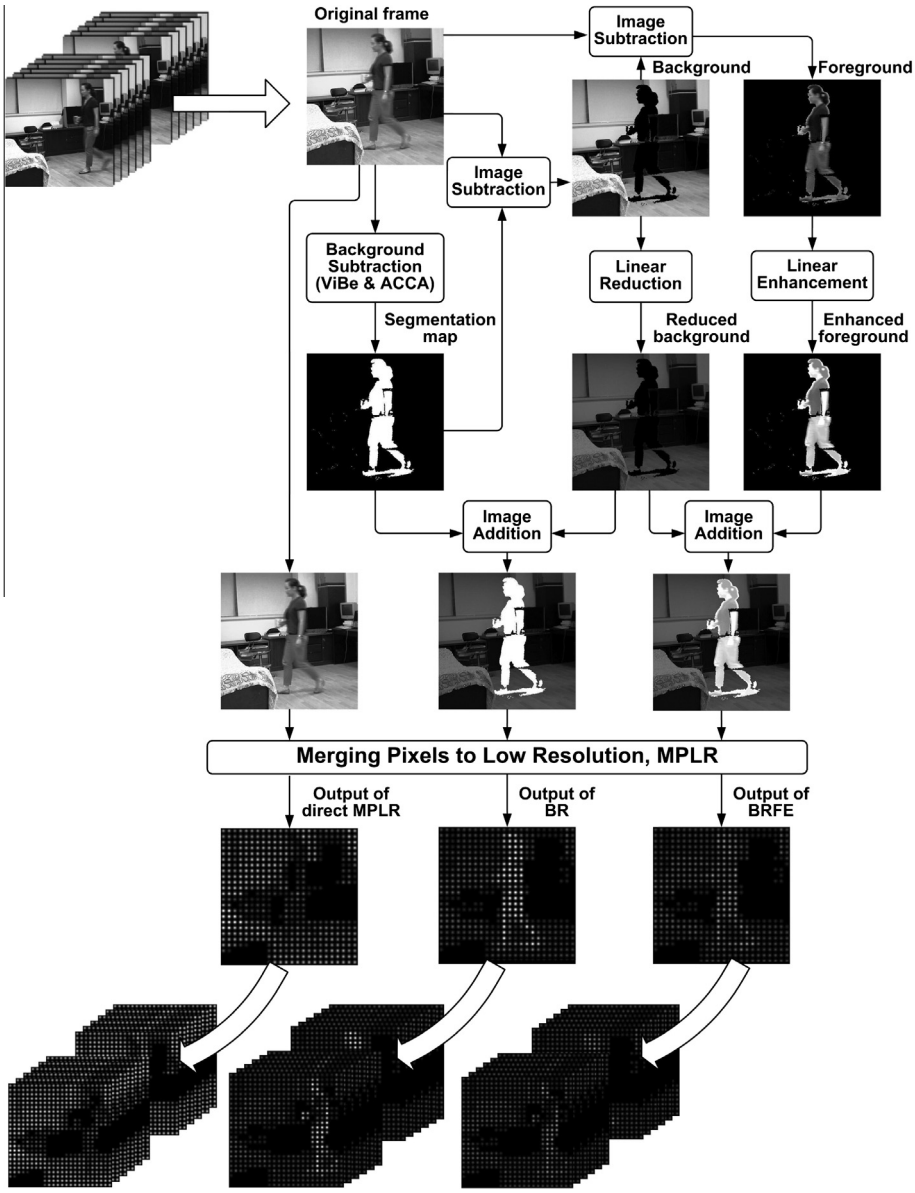

Fig. 2. Schematic diagram of the image-processing strategies. An image sample from the video "walking with a cup" is used to illustrate the step by step procedure.

$$P(t, t+dt) = e^{-\ln\left(\frac{N}{N-1}\right)dt} \tag{5}$$

Thus, the life span of a pixel in the background model is time-independent and decays exponentially. In addition, a random time subsampling factor ϕ, which selects new samples from an established background $f_B^k(x, y)$, is used for updating the background model over time. The random sampling mechanism propagates background pixel samples to guarantee spatial consistency and to update the background pixel model hidden by the foreground.

In this paper, parameters in ViBe were set to the same values used previously [3], i.e. $N = 20$, $R = 20$, $\#_{\min} = 2$, $\phi = 16$. The results corresponded to segmentation maps in which foreground areas were assigned to be the brightest.

2.2.1.2. Post-processing with adaptive connected component analysis. ViBe outperforms other background subtraction techniques, but it cannot completely solve some challenges, such as changes in illumination, and dynamic background and foreground aperture [37]. These technical challenges result in some incorrectly classified pixels in the ViBe segmentation maps. Incorrectly segmented regions will influence the final visual presentation, which could lead to confused percepts of moving objects. A post-processing method to deal with misclassification in each ViBe segmentation map was used before

the BS-based image strategies were presented. Common post-processing methods used morphology filtering alone or combined it with connected component grouping [26,30,35]. Based on the observed segmentation maps, in which the foreground was much bigger than the incorrectly classified spots that should have remained as background [16], we used an adaptive connected component analysis (ACCA) with median filtering to optimize the ViBe segmentation maps.

Firstly, the image is prefiltered by the median filter, which reduces the searching time for the connected components during ACCA processing. Then ACCA eliminates relatively bigger "foreground" regions using an adaptive threshold value. Unlike post-processing by Heikkila et al. [26], the ACCA filtering threshold is not a predefined constant, but is adaptively changed according to each segmentation map. Suppose c_i^k is a connected component in a 4-point neighborhood in a segmentation map at time k and $Area(c_i^k)$ is the number of pixels included in c_i^k, the general structure of the ACCA is described in Flowchart 1. The number of loops (i.e. n) is set to 2, which is optimal when considering computational speed and filtering results. Lastly, morphological region filling is used to repair the aperture because of the potential for removing small foreground pixels.

Flowchart 1. The general structure of adaptive connected component analysis in the post-processing.

begin
  Let n be the loop counter
  repeat
    Mark all connected component c_i^k in the input kth segmentation map
    Calculate the area of each c_i^k, denoted as $Area(c_i^k)$
    Sort all the $Area(c_i^k)$ and find the median as threshold T^k. Taking the larger if area number is even
    Set the pixel grayscale to 0 in c_i^k that $Area(c_i^k)$ is smaller than T^k
    Decreasing n
    until $n = 0$ stop and output the processed segmentation map
end

2.2.2. Background reduction

A background reduction (BR) image processing strategy based on the BS results was introduced to enhance foreground/background contrast. It is assumed that the difference between the post-processing segmentation map and the original grayscale image is the background $f_B^k(x,y)$. BR linearly decreases the gray level of $f_B^k(x,y)$ to half its value while maintaining the segmented foreground of the post-processing. The transformed background image $g_B^k(x,y)$ is expressed as

$$g_B^k(x,y) = T[f_B^k(x,y)] = \alpha f_B^k(x,y) + \beta \tag{6}$$

where α and β are constants set to 1/2 and 0 respectively. Then the transformed background image $g_B^k(x,y)$ is added to the post-processing segmentation map, to build a background-reduced video image.

2.2.3. Background reduction and foreground enhancement

Similar to the BR strategy, background reduction and foreground enhancement (BRFE) is also based on post-processing segmentation maps. Unlike BR, BRFE not only reduces the background, but also retains some grayscale information in the foreground. The foreground image $f_F^k(x,y)$ is generated by subtraction of the original frame and background $f_B^k(x,y)$ and is altered towards higher gray levels. The background is altered to lower gray levels ranging from 0 to 127 as in the BR strategy. Then the two are combined as the transformed image $g^k(x,y)$.

$$\begin{cases} g_F^k(x,y) = T[f_F^k(x,y)] = \alpha_1 f_F^k(x,y) + \beta_1 \quad \text{if } g_F^k(x,y) > 255, \quad g_F^k(x,y) = 255 \\ g_B^k(x,y) = T[f_B^k(x,y)] = \alpha_2 f_B^k(x,y) + \beta_2 \\ g^k(x,y) = g_F^k(x,y) + g_B^k(x,y) \end{cases} \tag{7}$$

We used $\alpha_1 = 1$, $\alpha_2 = 1/2$, and $\beta_1 = 128$, $\beta_2 = 0$, so that the foreground grayscale ranged from 128 to 255, while the darkened background ranged from 0 to 127.

2.2.4. Phosphene simulation

Pixel number of video image must be down-sampled to match the limited number of stimulating electrodes in a retinal implant. Taking $g^k(x,y)$ for example, $N \times N$ pixels with center coordinates (μ_x, μ_y) are merged into single pixel with uniform luminance value, which corresponds to the mean grayscale value of the original matrix, and is denoted as $A(\mu_x, \mu_y)$. Bicubic interpolation was used to avoid excessive distortion. Then a two-dimensional circular Gaussian distribution, $G(x,y)$ was applied to each pixel of the image, and this formed the Gaussian-like phosphenes [24]. The computation process is as follows:

$$g^k(x,y) = A(\mu_x, \mu_y) \cdot G(x,y) \tag{8}$$

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma^2}} \tag{9}$$

where σ represents the standard deviation of the particular Gaussian function around its horizontal and vertical center coordinates μ_x and μ_y, which is equal to the Gaussian width. The diameter of each Gaussian patch was set to 0.625° (32 × 32) or ~0.833° (24 × 24), corresponding to a simulated 20° × 20° prosthetic visual field such as used in an epiretinal prosthesis, in which the size of perceived phosphenes ranged from 0.4° to 2° [27]. Setting 8 grayscales appeared appropriate according to number of luminance levels attainable in human trials [14]. After MPLR pixelization, the original grayscale sequences or processed video images by BR and BRFE were converted and consisted of two different phosphene resolutions.

2.3. Subjects

Eighteen volunteers (9 female, 9 male) from Shanghai Jiao Tong University (aged between 20 and 25 years) with normal or corrected-to-normal visual acuity (20/20) participated. Before the formal experiment, subjects were instructed as to the purpose and experimental procedure, and then gave signed consent. Experiments were conducted in accordance with the Declaration of Helsinki.

2.4. Apparatus

Subjects were seated 50 cm in front of a 17″ CRT monitor (Proview Technology Co. WuHan, China, 800 × 600 resolution per display, 300 cd/m^2, 26° diagonal visual field) connected to a computer by a VGA distributor. In-house developed software to display experimental materials was written in C++ language and run in Windows 7 OS (Microsoft, Redmond, WA). Subject responses were recorded via a recording system. A high-speed video eye position tracking system (220 fps USB, Arrington Research Inc., Arizona, USA) was used to locate eye gaze and to control for eye movement.

2.5. Experimental procedure

Experiments were carried out monocularly (dominant eye) using central vision. Subjects were given a training and familiarization program before the formal experiment, during which they viewed six pixelated video clips of two different dynamic scenarios (1 indoor and 1 outdoor; not included in the experimental database). The subjects were shown a questionnaire before training and then required to describe in detail what they saw while viewing the video (e.g. "what is the moving object? What it is doing? Under what circumstance/location was the video taken? What objects are motionless?").

Formal experiments were divided into two sessions according to different video scenarios. In each session, subjects were shown two groups of 30 videos processed by three different strategies and at two resolutions (24 × 24 or 32 × 32). Presentation order of the strategies was counterbalanced across three subject groups (Table 1) in order to even out practice effect. The subjects were encouraged to describe what they had seen in as much detail as possible. Videos were automatically repeated three times (~30 s/video). There were four short breaks to reduce subject fatigue and the whole procedure lasted ~1 h.

2.6. Data analysis

Experimental performance was evaluated based on recognition accuracy (RA) according to separate descriptions of foreground and background information. A RA score was assessed by comparing the answers and descriptions against a set of standard answers given by a group of assessors (not subjects) who viewed the original videos. The more precise the subject's answers were compared with the standard answers, the higher the RA score was. For example, in the 'greeting' scene, if the subject recognized the presence of a person, they were given a basic score of one for foreground recognition; identifying the person's(s') movement (walking and hand waving) was given a score of one. A more precise description (e.g. walking and waving a hand with a handbag) was given a full score for foreground information. A correct description of the background was given a score of one, and additional information received additional points. The overall score for each condition was

Table 1
Order of test conditions.

Subject group (no. of subjects)	1st strategy tested	2nd strategy tested	3rd strategy tested
A (6)	MPLR	BR	BRFE
B (6)	BR	BRFE	MPLR
C (6)	BRFE	MPLR	BR

normalized to a percentage. In addition, the response time for correctly identifying a moving object was derived as the interval between the onset of object movement(s) and supplying the correct description.

SPSS for Windows (2010, IBM SPSS Inc. New York, USA) was used for statistical analysis. Results are expressed as mean ± SD. Statistically significant effects of resolution and image processing strategies were determined by a two-factor variance analysis (ANOVA), followed by a Bonferroni correction for multiple comparisons. Paired student's *t*-tests were used to determine statistically significant differences in recognition performance between different resolutions.

3. Results

3.1. Results of ViBe with post-processing with ACCA

Data from two experiment sequences were compared with the ground-truth and ViBe maps to evaluate the performance of ViBe with ACCA post-processing ("combined method").

Although most of the moving foreground was extracted from the image, the ViBe segmentation maps contained incorrectly classified regions due to factors such as a change in lighting conditions or waving trees. In Fig. 3, it is clear that some pixels were incorrectly labeled as foreground in the ViBe segmentation maps, but were rectified by post-processing with ACCA and appeared as background. In addition, apertures in the foreground were supplemented in post-processing results. The post-processing adaptively removed some small regions and filled up the foreground areas to some extent in the ViBe maps, thus, reducing inaccuracies in the ViBe segmentation. However, the combined method was unable to remove all mistaken pixels or regions due to the time-cost in ACCA looping. Consequently, some misclassified areas remained in the segmentation maps.

3.2. Experimental results of moving object recognition

3.2.1. Indoor scenes

Fig. 4 and Table 2 show the RA vs. image processing strategy for foreground and background recognition of indoor scenes at two resolutions. The highest scores were obtained after BRFE processing with a 32×32 phosphene array. RA values was close to that of the BRFE while implementing the BR strategy. The performances for all subjects declined with MPLR.

There was a statistically significant effect of both image processing strategy and resolution on RA scores ($F_{strategy} = 269.007$, $p < 0.001$; $F_{resolution} = 69.353$, $p < 0.001$) for recognition of foreground objects. But a significant interaction between them was not detected. *Post hoc* comparisons demonstrated that BR and BRFE significantly increased foreground RA scores compared with MPLR RA scores, whereas BR and BRFE RA scores were indistinguishable. Resolution also significantly affected performance for different image processing strategies (paired *t*-test, MPLR: $t = -3.220$, $p < 0.01$; BR: $t = -5.839$, $p < 0.001$; BRFE:

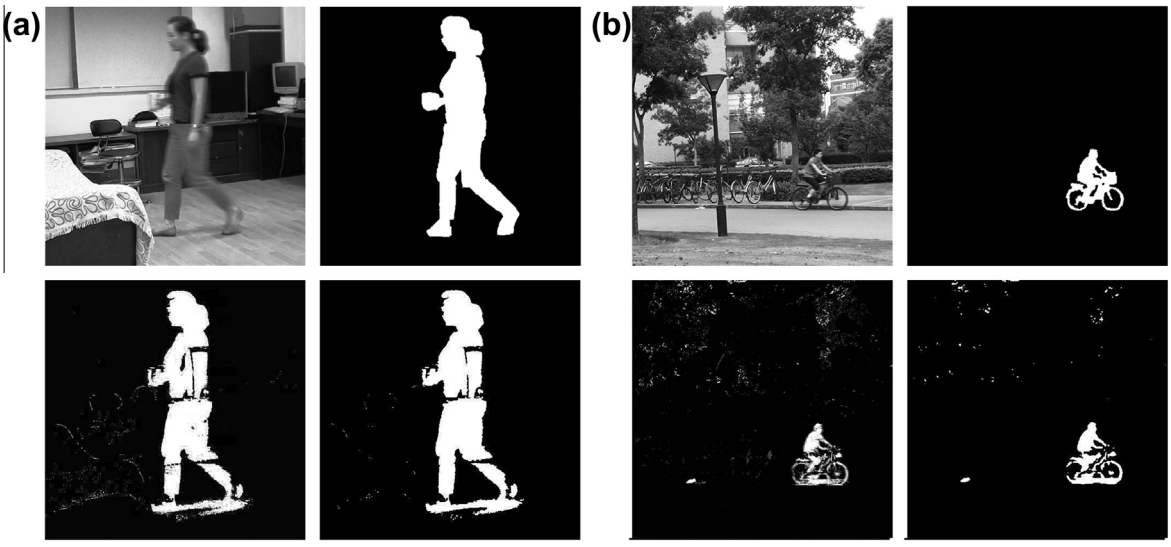

Fig. 3. Comparative foreground segmentation maps of ViBe and combined method taken from two sequences. (a) Indoor video "walking with a cup". (b) Outdoor video "bicycle riding". From left to right, the upper two images are the original frame and the ground-truth (manually labeled), respectively; the lower panels are the result of segmentation with ViBe (gray) or the combined method, respectively. The brightest regions are moving foreground; black areas represent the background scene.

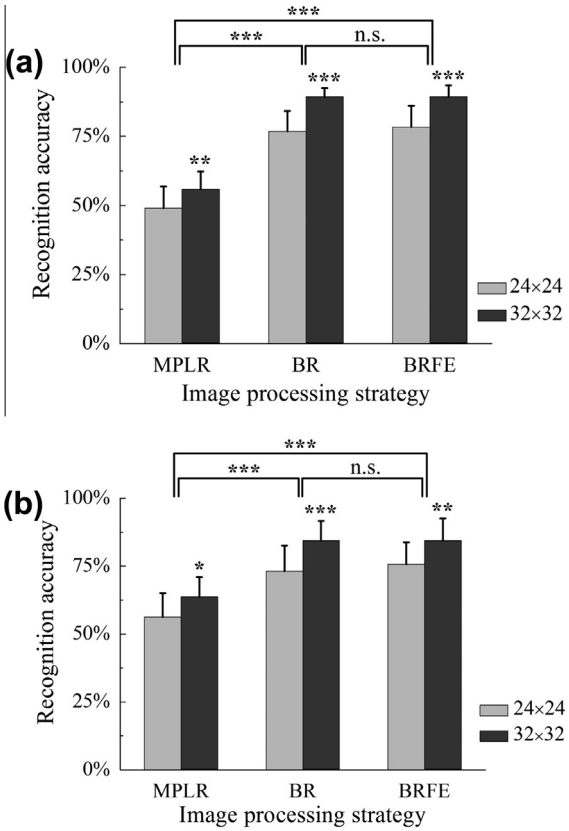

Fig. 4. Recognition accuracy of indoor scenes for two image resolutions and three different image processing strategies. (a) Foreground information. (b) Background information. (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, $n = 18$). Top lines show comparisons between strategies, notation above the bars indicate comparisons between resolutions.

Table 2
The recognition accuracy of indoor scenes by different image processing strategies at two resolutions.

Condition	Resolution	Recognition accuracy		
		MPLR	BR	BRFE
Foreground	24 × 24	49.06% ± 7.79%	76.88% ± 7.27%	78.44% ± 7.69%
	32 × 32	55.94% ± 6.38%	89.38% ± 3.10%	89.38% ± 4.03%
Background	24 × 24	56.25% ± 8.58%	73.13% ± 9.47%	75.63% ± 8.14%
	32 × 32	63.75% ± 7.19%	84.38% ± 7.27%	84.38% ± 8.14%

$t = -5.084$, $p < 0.001$). Similar to the low-resolution condition, performance differences, either between MPLR and BR or the BRFE condition, were significant. However, the difference between BR and BRFE was not significant.

Background RA scores based on BR and BRFE both increased by around 10% compared with MPLR scores (24 × 24 or 32 × 32 resolution). Image processing strategy and resolution had a significant impact on RA for the two resolutions ($F_{strategy} = 57.029$, $p < 0.001$, $F_{resolution} = 29.049$, $p < 0.001$), but there was not significant interaction between them. *Post hoc* comparisons showed that BR and BRFE significantly increased RA compared with MPLR, whereas the former two methods were indistinguishable. Resolution also significantly affected performance when using different image processing strategies (paired *t*-test, MPLR: $t = -2.236$, $p < 0.05$; BR: $t = -5.084$, $p < 0.001$; BRFE: $t = -3.217$, $p < 0.01$).

3.2.2. Outdoor scenes

Fig. 5 and Table 3 show the RA scores obtained after viewing outdoor scenes. The impact of the image processing strategy was highly significant, whether the task was recognizing the foreground ($F_{strategy} = 119.277$, $p < 0.001$) or the background ($F_{strategy} = 44.125$, $p < 0.001$). Resolution also played an important role in RA (foreground: $F_{resolution} = 29.347$, $p < 0.001$; background: $F_{resolution} = 10.696$, $p < 0.01$).

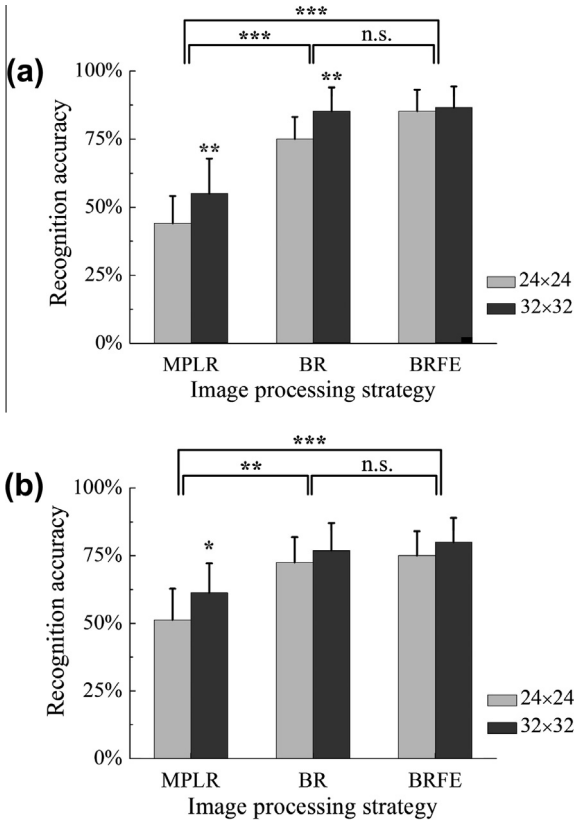

Fig. 5. Recognition accuracy of outdoor scenes after different image processing strategies at two resolutions. (a) Foreground. (b) Background. (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, $n = 18$). Top lines show comparisons between strategies, notation above the bars indicate comparisons between resolutions.

Table 3
The recognition accuracy of outdoor scenes by different image processing strategies at two resolutions.

Condition	Resolution	Recognition accuracy		
		MPLR	BR	BRFE
Foreground	24 × 24	44.06% ± 10.04%	75.00% ± 8.17%	85.31% ± 8.65%
	32 × 32	55.00% ± 12.90%	85.31% ± 8.65%	86.56% ± 7.69%
Background	24 × 24	51.25% ± 11.47%	72.50% ± 9.31%	75.00% ± 8.94%
	32 × 32	61.25% ± 10.88%	76.88% ± 10.15%	80.00% ± 8.94%

Table 4
The correct response times for different image processing strategies.

Resolution	MPLR	BR	BRFE
24 × 24	6.70 ± 17.91%	6.37 ± 18.66%	6.11 ± 17.63%
32 × 32	6.31 ± 20.99%	6.08 ± 20.26%	5.93 ± 18.89%

The mean RA for MPLR was rather low when subjects tried to recognize foreground objects at 24 × 24 resolution, in contrast to BR and BRFE, which greatly improved the RA. The foreground RA score with MPLR was lower than 60% when images were reduced to a 32 × 32 array; however, RA scores using BR and BRFE were significantly increased. Paired t-tests indicated that resolution significantly affected performance for the three image processing strategies (MPLR: $t = -3.487$, $p < 0.01$; BR: $t = -3.509$, $p < 0.01$; BRFE: $t = -5.745$, $p < 0.001$).

The RA for background objects significantly increased when BR was implemented with a low-resolution video image (24 × 24). RA scores increased even further with BRFE, which was ~15% higher compared with MPLR. However, differences between BR and BRFE were not significant. MPLR results were higher when using a 32 × 32 image compared with the lower resolution. Except for the BR and BRFE conditions, resolution did not have a significant impact on the RA under different image processing strategies.

The response times for correct recognition of a moving object were also analyzed at both resolutions (Table 4). The correct response times with BR and BRFE were shorter compared with MPLR, albeit not significantly ($p > 0.05$). The effect of resolution was also not significant ($p > 0.05$).

3.3. Processing speed of different image processing strategies

MPLR, BR and BRFE computational times were also compared with respect to for the presentation of achievable frames. Table 5 shows the average processing speed of all experiment sequences on our platform (2.93 GHz Core2 Duo CPU, 4 GB RAM, Red hat Linux 6.2, C implementation) and it is clear that MPLR had the fastest processing speed.

Down-sampling and decimation is an essential processing step in visual prosthesis prototypes. MPLR simulation is widely used in psychophysiological tests of visual function. BR and BRFE (implemented in the BS technique and post-processing) also utilized MPLR in last processing step, and thus are relatively more complex and slower. Although the average processing speed of the ViBe algorithm can achieve 200 frames/s [3], computational time still increased when it was integrated into image processing for prosthetic vision. BR based on Vibe reduced the grayscale of each input frame before MPLR and fulfilled the requirements for real-time processing even though its speed fell sharply to 32 frames/s. The computational speed of BRFE was slower than the former two algorithms due to an additional foreground-enhancement step in the processing.

4. Discussion

We propose two strategies to highlight moving objects based on a BS technique (ViBe) that can be used for danger avoidance in independent mobility and navigation by a prosthesis wearer. Subjects viewed dynamic scenes in the form of a limited number of discrete dots in a restricted field of view to simulate the percept of a prosthesis wearer. The results indicated that perception of moving objects in a pixelized visual presentation of a real-life scene was affected by the image processing strategy as well as resolution, and our background-subtraction-based strategies were advantageous to recognition of a dynamic scene in low-resolution prosthetic vision.

McCarthy and Barnes [32] applied a motion detection technique to detect and enhance a moving object in a frontal view of simulated vision. In agreement with them, our study advocate highlighting the detected moving object when using low-resolution prosthetic vision. Due to differences in the applied situation, the techniques used for movement detection in the two works differed. McCarthy and Barnes focused on the early perception of the incoming objects and tested the feasibility of a time-to-contact map by sequence analysis and image presentation. We utilized a universal background subtraction algorithm-ViBe and demonstrated the effectiveness of background-subtraction-based strategies to facilitate motion detection and enhance the profile of a moving object. Our strategies also have the ability to deal with the incoming object detection and enhancement, based on the video testing results of Vibe in [3]. Yet for avoiding the effect of resolution, each moving object in current experimental materials was in profile to keep a consistent viewing angle.

4.1. ViBe with post-processing with ACCA

An adaptive post-processing filtering method was used to deal with incorrect pixel classification in ViBe segmentation results, which effectively revised certain incorrectly assigned regions in the ViBe segmentation maps, and can be applied in other BS algorithms to optimize detected foreground images. Outcomes of the two strategies were largely determined by the quality of BS segmentation. Some challenges, such as changes in illumination, cannot be completely resolved by computer vision, and this influenced the validity of the ViBe algorithm and ACCA, which is not substantially solved by post-processing. The effectiveness of BR or BRFE was affected to a greater or lesser degree by the quality of the segmentation. Therefore, other effective background subtraction techniques need to be tested for further improvement of moving object extraction, such as Eigenbackground based statistical illumination proposed by Vosters et al. [41], in particular its handling of rapid changes in illumination. We have only investigated a static camera scenario to study the feasibility of increasing contrast when perceiving a moving object and are based on ViBe, which is more effective under static camera conditions. Using embedded motion sensors or other algorithmic techniques that deal with a moving camera could make the strategies more suitable for a mobile camera.

4.2. Effect of image processing strategy – MPLR, BR and BRFE

The MPLR had the lowest RA scores for foreground or background recognition, regardless of image resolution. Because of the low resolution it was difficult to correctly perceive an object in a complex scene, whether it belonged to the foreground

Table 5
Computational speed of different image processing strategies.

Strategy	MPLR	BR	BRFE
Speed (frames/s)	80	32	23

or the background, especially with insufficient luminance contrast. Thus the ability to recognize moving objects is significantly influenced under this condition.

Although the ROI-zooming technique appears to be effective for static object and scene recognition, it is more problematic for recognizing moving objects. The perceived information of moving objects, such as position and velocity are altered when ROI-zooming, which will influence the viewer's perception. According to behavioral and neurophysiological studies, visual attention is influenced by feature differences between the target and non-target, referred to as local feature contrast by Hans-Christoph [25]. Human attention is biased towards objects that are larger, brighter, and fast moving [38], and people always notice regions with higher contrast more easily. In light of this, the ViBe technique with ACCA post-processing was implemented to extract the moving foreground from the whole scene. Based on the BS-segmentation results, BR weakened background information. Our analysis demonstrated that this was not related to resolution. The ability to perceive moving objects in a dynamic scene increases substantially by increasing the grayscale contrast between a moving foreground and static background. Improved performance compared with the MPLR strategy also resulted in faster recognition of moving objects, allowing more time to perceive background information. This will be beneficial to users in the timely recognition of potential risks when involved in outside activities.

Gray levels have an impact on object and face recognition when using prosthetic vision [4,36], thus, the BRFE strategy was not only used to weaken the background, but also used to assign four higher grayscale levels to the distribution in foreground regions. Similar to the BR method, BRFE was significantly better than MPLR in helping the subject to perceive a dynamic low resolution scene. Owing to enhancement of the foreground information, BRFE was slightly better than BR in each experimental situation. This indicated that adding grayscale information was also beneficial in perceiving moving objects. However, proportionally enhancing the gray level did not significantly increase the RA scores, and may be related to the limited number of gray levels constrained by array resolution. If the number of electrodes capable of stimulating the retina is significantly increased and then more gray levels added, the BRFE image processing strategy could provide more details of the objects and scene, and further improve the perceptual capacity of prosthetic wearers. On the other hand, The results show that the strategies used had no significant effect on correct response time. Subjects were required to describe the scene in detail according to a prescribed list. We suggest that most subjects spent more time devoted to comprehension in order to get higher recognition scores, rather than respond as soon as possible. However, correct response times to BR and BRFE images were faster than those related to images made without a strategy condition.

The computational speed of each strategy and the two proposed methods can fulfill real-time processing requirements. However, processing will take longer when combined with other techniques necessary for a visual prosthesis. It is believed that future development of a high-level computing platform and algorithm optimization will make it possible for BS-based strategies to be applied to a prosthesis.

4.3. Effect of resolution

Nearly all indoor and outdoor scenes were significantly affected by resolution, indicating that resolution is an important factor for recognition of moving objects in a complex scene. This is in agreement with conclusions based on an evaluation of the minimal amount of visual information necessary for functional tasks such as visual acuity [9], reading [11,12,19,45], and mobility [10,20]. The number of pixels did not significantly enhance the recognition capacity of the background in outdoor scenes with BR and BRFE. However, even with a 24×24 resolution and BR and BRFE, both background RA scores were above 70%, which was a relatively good performance for background recognition. This in part may be due to the foreground information providing clues as to the background scene, such as trees that are usually beside by the road. A 32×32 resolution increased the average RA to 80%, indicating that providing more resolution improved performance to some extent, but did not totally compensate for much of the information lost due to pixelization.

4.4. Limitations

Although simulation provides a feasible way to approximate the experience of prosthesis wearers, it is still idealized relative to the actual visual percept achieved with an irregular retinotopic map [21,27,40] and/or phosphene dropout [7,27] such as reported by chronic implant recipients. These factors will be considered in future studies of movement perception. Because the ability to recognize objects and faces in low-quality images is related to edge quality [13,28], enhancing contour information based on BS techniques is a promising method whereby perception of moving objects can be enhanced for prosthetic vision.

5. Conclusions

Visual prosthetic research has made significant progress since the first human tests over half a century ago, and its focus has already shifted from one of feasibility to one of optimizing the visual presentation and technical development. We focused on exploring effective image processing strategies to optimize information content and to improve the perception of moving objects. We demonstrated that adaptive post-processing effectively improved foreground segmentation performance. Importantly, results of psychophysiological experiments indicated that BR and BRFE strategies were advantageous

to the perception of movement. The BRFE strategy added more gray scale information and subjects attained higher recognition accuracy. It is hoped that background-subtraction-based image processing strategies will encourage further development of image-processing modules for a visual prosthesis that will assist implant recipients to avoid dangerous situations and attain independent mobility in daily life.

Acknowledgements

The authors are grateful to the volunteers and thank Dr. T. FitzGibbon for comments on drafts of the manuscript. This research is supported by The National Basic Research Program of China (973 Program, 2011CB707503); The National Natural Science Foundation of China (61273368, 91120304, 81171377); National High Technology Research and Development Program of China (863 Program, 2009AA04Z326); The National Key Technology R&D Program (2007BAK27B04, 2008BAI65B03); The China Postdoctoral Science Foundation (2013M540363); Shanghai Municipal Physical Culture Bureau Scientific and Technological Project (11JT010); Shanghai Science and Technology Development Funding (10231204300); The 111 Project from the Ministry of Education of China (B08020).

References

[1] R.G. Abbott, L.R. Williams, Multiple target tracking with lazy background subtraction and connected components analysis, Mach. Vis. Appl. 20 (2) (2009) 93–101.
[2] O. Barnich, M.V. Droogenbroeck, ViBE: A powerful random technique to estimate the background in video sequences, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, Taipei 2009, pp. 945–948.
[3] O. Barnich, M.V. Droogenbroeck, ViBe: a universal background subtraction algorithm for video sequences, IEEE Trans. Image Process. 20 (6) (2011) 1709–1724.
[4] J. Boyle, W. Boles, A. Maeder, Challenges in digital imaging for artificial human vision, in: Proceedings of the SPIE, SPIE – The International Society for Optical Eng, San Jose, CA, USA, 2001, pp. 22–25.
[5] J. Boyle, A. Maeder, W. Boles, Static image simulation of electronic visual prostheses, in: Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001, 2001, pp. 85–88.
[6] J.R. Boyle, A.J. Maeder, W.W. Boles, Region-of-interest processing for electronic visual prostheses, J. Electron. Imag. 17 (1) (2008) 0130021–01300212.
[7] G.S. Brindley, The sensations produced by electrical stimulation of the visual cortex, J. Physiol.-Lond. 196 (2) (1968) 479–493.
[8] S. Brutzer, B. Hoferlin, G. Heidemann, Evaluation of background subtraction techniques for video surveillance, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR' 2011, Providence, RI, 2011, pp. 1937–1944.
[9] K. Cha, K. Horch, R. Normann, Simulation of a phosphene-based visual field: visual acuity in a pixelized vision system, Ann. Biomed. Eng. 20 (4) (1992) 439–449.
[10] K. Cha, K.W. Horch, R.A. Normann, Mobility performance with a pixelized vision system, Vis. Res. 32 (7) (1992) 1367–1372.
[11] K. Cha, K.W. Horch, R.A. Normann, D.K. Boman, Reading speed with a pixelized vision system, J. Opt. Soc. Am. A – Opt. Image Sci. Vis. 9 (5) (1992) 673–677.
[12] X. Chai, W. Yu, J. Wang, Y. Zhao, C. Cai, Q. Ren, Recognition of pixelized Chinese characters using simulated prosthetic vision, Artif. Organs 31 (3) (2007) 175–182.
[13] M.H. Chang, H.S. Kim, J.H. Shin, K.S. Park, Facial identification in very low-resolution images simulating prosthetic vision, J. Neur. Eng. 9 (4) (2012) 046012.
[14] S.C. Chen, G.J. Suaning, J.W. Morley, N.H. Lovell, Simulating prosthetic vision: I. Visual models of phosphenes, Vis. Res. 49 (12) (2009) 1493–1506.
[15] S.C. Chen, G.J. Suaning, J.W. Morley, N.H. Lovell, Simulating prosthetic vision: II. Measuring functional capacity, Vis. Res. 49 (19) (2009) 2329–2343.
[16] S.-C.S. Cheung, C. Kamath, Robust techniques for background subtraction in urban traffic video, in: S. Panchanathan, B. Vasudev (Eds.), Visual Communications and Image Processing 2004, SPIE, San Jose, CA, USA, 2004, pp. 881–892.
[17] G. Dagnelie, Psychophysical evaluation for visual prosthesis, Ann. Rev. Biomed. Eng. 10 (2008) 339–368.
[18] G. Dagnelie, Visual prosthetics 2006: assessment and expectations, Exp. Rev. Med. Dev. 3 (3) (2006) 315–325.
[19] G. Dagnelie, D. Barnett, M.S. Humayun, R.W. Thompson, Paragraph text reading using a pixelized prosthetic vision simulator: parameter dependence and task learning in free-viewing conditions, Invest. Ophthal. Vis. Sci. 47 (3) (2006) 1241–1250.
[20] G. Dagnelie, P. Keane, V. Narla, L. Yang, J. Weiland, M. Humayun, Real and virtual mobility performance in simulated prosthetic vision, J. Neur. Eng. 4 (1) (2007) 92–101.
[21] J. Delbeke, M. Oozeer, C. Veraart, Position, size and luminosity of phosphenes generated by direct optic nerve stimulation, Vis. Res. 43 (9) (2003) 1091–1102.
[22] J.A. Dowling, A.J. Maeder, W.W. Boles, Intelligent image processing constraints for blind mobility facilitated through artificial vision, in: 8th Australian and New Zealand Intelligent Information Systems Conference, Queensland University of Technology, 2003.
[23] C.D. Eiber, N.H. Lovell, G.J. Suaning, Attaining higher resolution visual prosthetics: a review of the factors and limitations, J. Neur. Eng. 10 (1) (2013) 011002.
[24] A.P. Fornos, J. Sommerhalder, B. Rappaz, A.B. Safran, M. Pelizzone, Simulation of artificial vision, III: do the spatial or temporal characteristics of stimulus pixelization really matter?, Invest Ophthal. Vis. Sci. 46 (10) (2005) 3906–3912.
[25] N. Hans-Christoph, The role of features in preattentive vision: comparison of orientation, motion and color cues, Vis. Res. 33 (14) (1993) 1937–1958.
[26] J. Heìikkilä, O. Silven, A real-time system for monitoring of cyclists and pedestrians, in: Second IEEE Workshop on Visual Surveillance, 1999, VS'99, Fort Collins, CO, 1999, pp. 74–81.
[27] M.S. Humayun, J.D. Weiland, G.Y. Fujii, R. Greenberg, R. Williamson, J. Little, B. Mech, V. Cimmarusti, G. Van Boemel, G. Dagnelie, E. de Juan Jr, Visual perception in a blind subject with a chronic microelectronic retinal prosthesis, Vis. Res. 43 (24) (2003) 2573–2581.
[28] B. Justin, M. Anthony, B. Wageeh, Inherent visual information for low quality image presentation, in: APRS Workshop on Digital Image Computing (WDIC) : Medical Applications of Image Analysis, 2003, pp. 51–56.
[29] P. Kaewtrakulpong, R. Bowden, An improved adaptive background mixture model for realtime tracking with shadow detection, in: Proceedings of the European Workshop Advances in Video Based Surveillance Systems, London, UK, 2001.
[30] C.C.-P. Li, L. Kurz, A new approach to the detection of moving objects, Inform. Sci. 103 (1) (1997) 115–134.
[31] B.P.L. Lo, S.A. Velastin, Automatic congestion detection system for underground platforms, in: Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2001, Hong Kong, 2001, pp. 158–161.
[32] C. McCarthy, N. Barnes, Time-to-contact maps for navigation with a low resolution visual prosthesis, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), 2012, pp. 2780–2783.
[33] J. Sommerhalder, E. Oueghlani, M. Bagnoud, U. Leonards, A.B. Safran, M. Pelizzone, Simulation of artificial vision: I. Eccentric reading of isolated words, and perceptual learning, Vis. Res. 43 (3) (2003) 269–283.

[34] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999, Fort Collins, CO, 1999, pp. 246–252.
[35] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. Patt. Anal. Mach. Intell. 22 (8) (2000) 747–757.
[36] R.W. Thompson, G.D. Barnett, M.S. Humayun, G. Dagnelie, Facial recognition using simulated prosthetic pixelized vision, Invest. Ophthal. Vis. Sci. 44 (11) (2003) 5035–5042.
[37] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: principles and practice of background maintenance, in: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, Kerkyra, 1999, pp. 255–261.
[38] A. Treisman, S. Gormican, Feature analysis in early vision: evidence from search asymmetries, Psychol. Rev. 95 (1) (1988) 15–48.
[39] J.J. van Rheede, C. Kennard, S.L. Hicks, Simulating prosthetic vision: optimizing the information content of a limited visual display, J. Vis. 10 (14) (2010) 1–14.
[40] C. Veraart, C. Raftopoulos, J.T. Mortimer, J. Delbeke, D. Pins, G. Michaux, A. Vanlierde, S. Parrini, M.-C. Wanet-Defalque, Visual sensations produced by optic nerve stimulation using an implanted self-sizing spiral cuff electrode, Brain Res. 813 (1) (1998) 181–186.
[41] L. Vosters, C. Shan, T. Gritti, Real-time robust background subtraction under rapidly changing illumination conditions, Image Vis. Comput. 30 (12) (2012) 1004–1015.
[42] J.D. Weiland, M.S. Humayun, Visual prosthesis, Proc. IEEE 96 (7) (2008) 1076–1084.
[43] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. Patt. Anal. Mach. Intell. 19 (7) (1997) 780–785.
[44] Y. Zhao, Y. Lu, Y. Tian, L. Li, Q. Ren, X. Chai, Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision, Inform. Sci. 180 (16) (2011) 2915–2924.
[45] Y. Zhao, Y. Lu, J. Zhao, K. Wang, Q. Ren, K. Wu, X. Chai, Reading pixelized paragraphs of Chinese characters using simulated prosthetic vision, Invest. Ophthal. Vis. Sci. 52 (8) (2011) 5987–5994.
[46] D.D. Zhou, J.D. Dorn, R.J. Greenberg, The Argus@ II retinal prosthesis system: an overview, in: IEEE International Conference on Multimedia and Expo Workshops, IEEE, San Jose, 2013, pp. 1–6.
[47] E. Zrenner, Will retinal implants restore vision?, Science 295 (5557) (2002) 1022–1025
[48] E. Zrenner, D. Besch, K.U. Bartz-Schmidt, F. Gekeler, V.P. Gabel, C. Kuttenkeuler, H. Sachs, H. Sailer, B. Wilhelm, R. Wilke, Subretinal chronic multi-electrode arrays implanted in blind patients, Invest. Ophthal. Vis. Sci. 47 (5) (2006) 1538.