J. Neural Eng. **7** (2010) 016006 (10pp) doi:10.1088/1741-2560/7/1/016006

Saliency-based image processing for retinal prostheses

N Parikh[1], **L Itti**[2] **and J Weiland**[3]

[1] Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA
[2] Department of Computer Science, Psychology and Neuroscience, University of Southern California, Los Angeles, CA, USA
[3] Departments of Ophthalmology and Biomedical Engineering, University of Southern California, Los Angeles, CA, USA

E-mail: njparikh@usc.edu, itti@usc.edu and jweiland@usc.edu

Received 31 August 2009
Accepted for publication 7 December 2009
Published 14 January 2010
Online at stacks.iop.org/JNE/7/016006

Abstract

We present a computationally efficient model for detecting salient regions in an image frame. The model when implemented on a portable, wearable system can be used in conjunction with a retinal prosthesis, to identify important objects that a retinal prosthesis patient may not be able to see due to implant limitations. The model is based on an earlier saliency detection model but has a reduced number of parallel streams. Results of a comparison between the areas detected as salient by the algorithm and areas gazed at by human subjects in a set of images show a correspondence which is greater than what would be expected by chance. Initial results for a comparison of the execution speed of the two algorithm models for each frame on the TMS320 DM642 Texas Instruments Digital Signal Processor suggest that the proposed model is approximately ten times faster than the original saliency model.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

An electronic retinal prosthesis is under development, to treat blinding diseases like retinitis pigmentosa (RP) and age-related macular degeneration (AMD) [1]. In RP and AMD, the photoreceptor cells are affected while other retinal cells remain relatively intact. Photoreceptor cells convert light information entering the retina into electrical signals and hence the progressive loss of these cells leads to a gradual loss of vision in patients. The retinal prosthesis aims to provide partial vision by electrically activating the remaining cells of the retina. Current retinal prosthesis prototypes use external components to acquire and code image data for transmission to an implanted retinal stimulator. The external system consists of a small camera to capture video in real time and a portable video processor to convert image data to a series of command signals which are wirelessly transmitted to the implanted retinal stimulator.

Human monocular vision has a field of view close to 160° [2]. Due to surgical limitations on implant size, current retinal prostheses only stimulate the central 15–20° field of view. Prototype systems range from 16 to 1550 electrodes [3–6], which is well below the resolution of the retina in this region, even if every electrode can create an independent pixel. If the entire camera image (between 40° and 60° field of view) is compressed to fit the central field, there will be a loss of resolution and miniaturization of objects, with a likely decrease in the quality of vision. Whereas, if only the central 15–20° field of view from the camera image is extracted and stimulated electrically, the visual information will be more organized and perceivable to the recipients. However, peripheral information will be lost, severely hampering mobility. Hence, there is a need for a specific image processing algorithm which could be used to overcome the loss of peripheral information due to the limited field of view.

Retinal prosthesis research can involve image processing in a number of different ways. Several studies have conducted simulated vision experiments with normal sighted volunteers performing reading or mobility tasks. The goal of these experiments was to test visual task performance as a function of pixel number, density and quality. These have been recently

reviewed [7] and collectively, these studies suggest that 600–1000 electrodes are needed for functional vision. Another facet of image processing research related to retinal prostheses involves the conversion of the image data into a stimulus pattern that best conveys the desired visual perception. Asher *et al* [8] propose real-time image processing algorithms and transformations to simulate the different functions of the various cell layers and cell layer connections in the retina. They also propose a conversion method for transforming the visual information into electrical current patterns. Hallum *et al* [9] also propose a method for converting an image frame captured by the camera into low resolution modulated charge injections. Finally, image processing has been proposed to enhance certain features of an image that may be important to the user. Boyle *et al* [10] examined accentuating certain image features. One finding from this study suggested the utility of important maps, like the ones created by the bottom-up visual attention saliency model by Itti *et al* [11–15]. With this in mind, we propose an image processing algorithm based on the saliency detection model by Itti *et al* to find the important and salient regions in the entire image frame and cue subjects toward the direction of the salient region.

The retina along with higher visual processes guides visual attention in the visual cortex. Visual attention binds information from multiple parallel processes carrying motion, depth, color and form information in the visual cortex [16]. During visual search different information from these processes is combined first, and the output guides the attention deployment process [17, 18]. Computational models based on saliency detection have been used in computer vision and robotics to predict important areas in the visual field. A first model based on the feature integration theory was proposed by Koch and Ullman [19]. The model is a bottom-up saliency detection model that computes a saliency map by combining several basic features which undergo parallel processing. Based on this model, a bottom-up model based on primate vision was proposed by one of the authors of this paper [11–15]. This model (hereon referred to as the 'full model') forms the basis of many implementations of visual attention in robotics and artificial intelligence [20] and also forms the basis of the work proposed here. We propose an algorithm (hereon referred to as the 'new model') that is based on the full model, but with simplifications to increase efficiency, to allow execution on a portable processor. In this paper, we describe the new model in detail and verify that it can predict human gaze using a library of images and human observers.

2. Methods

2.1. New model algorithm

The new model (figure 1) uses three information streams: color saturation, intensity and edge information. These information streams are extracted by converting the input image from the RGB color space to the HSI (hue–saturation–intensity) color space. This conversion can be done in various ways. The conversion for our algorithm was done using the function rgb2hsv in Matlab from Mathworks Inc. Nine scales of dyadic Gaussian pyramids [21] are created for the saturation (S), intensity (I) and edge (E) information by successively low pass filtering and down sampling by a factor of 2. Edge pyramids are created from the intensity stream based on Laplacian pyramid generation [22, 23]. For each level of the pyramid, the edge pyramid image is created as a point-by-point subtraction between the intensity image at that level and the interpolated intensity image from the next level.

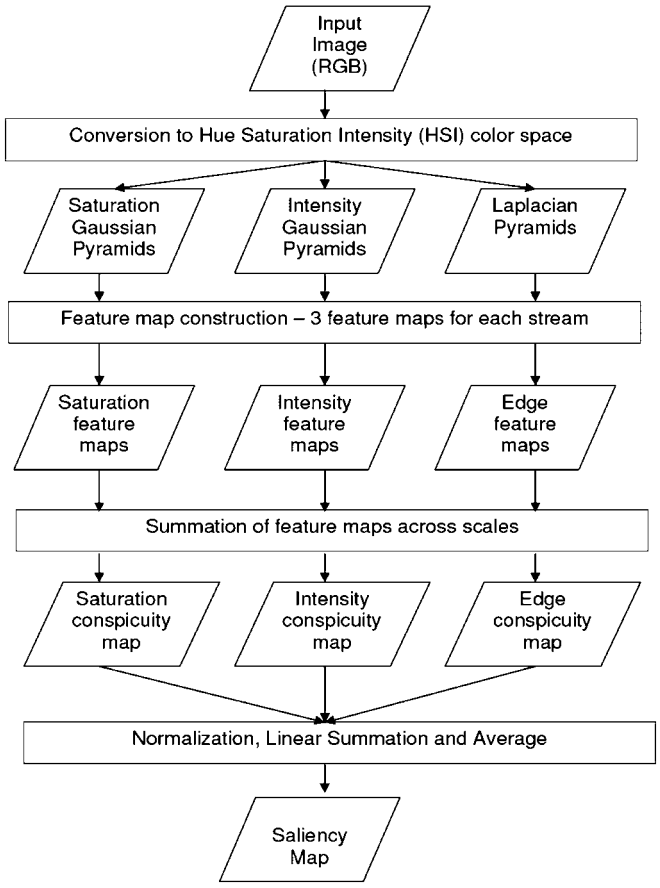

Figure 1. Diagram of the proposed algorithm.

Center–surround mechanisms observed in the visual receptive fields of the primate retina are then implemented computationally to create feature maps for each information stream. Center–surround interactions are modeled as the difference between the coarse and fine scales of the pyramids [11–14]. Feature maps are created from only four scales with the center scales 'c' at levels (3, 4) and surround scales 's' at levels (6, 7) where the original image is at level 0 of the pyramid. For $c \in \{3, 4\}$ and $s = c + \delta$ where $\delta \in \{3, 4\}$ and $s < 8$, a set of three feature maps is created for each stream. Point-by-point subtraction between the values of the pyramids at the finer and coarser scales is carried out after interpolating the coarser scale to the finer scale using bilinear interpolation. Absolute values of the subtraction are calculated for the saturation and intensity streams and all feature-maps are decimated by dropping the appropriate number of pixels to be 1/16th the size of the original image:

$$S(c, s) = |S(c) - S(s)| \tag{1}$$

$$I(c, s) = |I(c) - I(s)| \tag{2}$$

$$E(c, s) = E(c) - E(s). \tag{3}$$

A linear summation across these feature maps for the different information streams forms the conspicuity maps—S_c for color saturation, I_c for intensity and E_c for edge information:

$$S_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} S(c, s) \tag{4}$$

$$I_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} I(c, s) \tag{5}$$

$$E_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} E(c, s). \tag{6}$$

The conspicuity maps undergo normalization [11–15] referred to by the operator $\mathcal{N}$ in the equations. Normalization is an iterative process that promotes maps with a small number of peaks with strong activity and suppresses maps with many peaks of similar activity. Each conspicuity map is first normalized to a fixed range between 0 and 1. Thereafter, a two-dimensional difference of Gaussian filter (DoG) is convolved with the map iteratively. The output is summed with the original map and negative values are set to zero. The DoG filter results in the excitation of each pixel with inhibition from neighboring pixels. The DoG filter function is calculated as stated below:

$$\mathrm{DoG}(x, y) = \frac{0.5\,\mathrm{e}^{-(x^2+y^2)/(2\sigma_{\mathrm{ex}})}}{2\Pi\sigma_{\mathrm{ex}}^2} - \frac{1.5\,\mathrm{e}^{-(x^2+y^2)/(2\sigma_{\mathrm{inh}})}}{2\Pi\sigma_{\mathrm{inh}}^2}, \tag{7}$$

where $\sigma_{\mathrm{ex}} = 2\%$ and $\sigma_{\mathrm{inh}} = 25\%$ of the input image width.

Intensity and saturation conspicuity maps use three normalization iterations, and edge conspicuity maps use one normalization iteration. The number of iterations for normalization is chosen based on the computational load and pilot studies that examined different iterations of normalization and their effects on the maps. The three normalized conspicuity maps are linearly summed and their average forms the final salience map which again undergoes a three-iteration normalization. The region around the pixel with the highest grayscale value in the final salience map signifies the most salient region:

$$S = \mathcal{N}\left(\frac{\mathcal{N}(S_c) + \mathcal{N}(I_c) + \mathcal{N}(E_c)}{3}\right). \tag{8}$$

There are a few key differences between the two models which make the new model less computationally intensive compared to the full saliency model. The new model uses only 3 information streams for processing (versus 7 in the full model), 4 scales of Gaussian pyramids (versus 6), 18 feature maps (versus 42). Instead of using the two color opponent streams as found in the primate retina, the new model uses color saturation. Color saturation information will indicate purer hues with higher grayscale values and impure hues with lower grayscale values. One stream of edge information is used instead of four different orientation streams. For creating feature maps, the new model focuses on the coarser scales for center and surround which represent low spatial frequency information in the image.

2.2. DSP implementation

The retinal prosthesis image processing system is designed to be a portable module worn on the body. For this reason, the module should be lightweight and compact with low power requirements so that it can operate for several hours on a small battery. Low power requirements may restrict the amount of computation that can be carried out by the image processing unit. With this in mind, for research purposes, the algorithm has been implemented on a Texas Instruments Imaging Developers Kit (IDK) TMS320 DM642 [24]. The DM642 chip is a 720 MHz fixed-point processor and the IDK is specifically designed to aid the development of image processing algorithms. The IDK is not a portable board (it includes many functions) but in general, DSPs are designed for low-power, portable applications, so the technological path is clear.

As a first step toward analyzing the computational speed of the new model, we implement the new model and full model (partial) on the DSP-IDK. The algorithms are modeled in Simulink from Mathworks Inc., and then ported to the DSP. For efficiency, filtering has been implemented using separable one-dimensional filters for both the models. Only the intensity stream of the full model is implemented (one of the seven streams of the full model). Fixed-point hardware can implement numbers in both fixed-point precision and floating-point precision. Both the algorithm implementations are in single-precision floating-point format and are not optimized.

2.3. Model validation using gaze data

Gaze experiments were carried out with five human subjects after the approval of the Institutional Review Board at the University of Southern California. A signed informed consent was obtained from each participant of the study. Subjects were required to have English speaking and reading knowledge, be 18+ years of age, not have a history of vertigo, motion sickness or claustrophobia; cognitive or language/hearing impairments, and have a visual acuity of 20/40 or better with normal or corrected vision (with lenses). Visual acuity testing was carried out in the lab using a Snellen visual acuity eye chart.

Gaze data were acquired using an eye tracking system from Arrington Research, Inc., Scottsdale, AZ. This system consists of a Z800 3D Visor Head Mounted Display (HMD) with a diagonal field of view of 40°. Images on the HMD are displayed at a resolution of 800 × 600 pixels. The Viewpoint eye tracking software from Arrington Research recorded data at a frequency of 60 Hz using pupil tracking. Subjects were seated at a table with their head rested on a chin rest. A 12 point rectangular grid calibration process was used. Subjects were asked to look at the center of 12 different squares that would successively appear on the HMD screen. After this, as a measure of calibration, a test image consisting of a circle in the center of the screen was shown and subjects were asked to look at the center of the circle. Recording was not done until good calibration was obtained. Good calibration is defined as a rectangular grid mapped from the gaze points of the subjects when looking at the 12 squares. A set of 150 natural images

was displayed on the HMD with each image being shown for 3 s. The images consisted of outdoor and indoor environments. Subjects were instructed to freely gaze at the image. To avoid biasing the subjects, no other instructions were given. Between images, the test circle image was displayed for subjects to rest their eyes. However, after every three images, subjects were instructed to look at the center of the circle and the calibration at this point was noted. This helped to keep track of any calibration drift during the experiments. In post-processing, the recorded gaze point data were then corrected by any offset to get the new gaze point data, corrected for calibration drift.

The collected gaze data were filtered for fixations and saccades using custom fixation and saccade filtering software freely available on http://ilab.usc.edu as part of the Neuromorphic Vision Toolkit. Data were analyzed using gaze fixation points from the data set. Fixation data points may not account for drifts in eye movements. However, by taking a circular aperture around each fixation point during data analysis, effects of drifts as well as slight calibration offsets can be avoided.

Analysis of gaze data was carried out using methods used by Itti [25] and Peters *et al* [26] to analyze the contribution of bottom-up saliency to human eye movements. For each image, gaze data points from all subjects were pooled together for analysis. For the same set of 150 input images, salience maps created by the full model were used for a comparison of results with salience maps created by the new model.

2.3.1. Analysis with the ratio of medians method [25]. S_h is defined as the highest value of saliency within a circular aperture of diameter 5.6° centered at the fixation point. High values of S_h indicate that the human observer fixated at a highly salient region.

S_r is defined as the highest value of saliency within a 5.6° circular aperture, centered at a random point chosen from a uniform distribution.

S_{max} is defined as the maximum value of saliency in the salience map of the image.

Each image will have approximately between 20 and 40 gaze points after combining gaze data from all subjects. The same number of points are randomly chosen from a uniform distribution to calculate S_r. To get a more accurate estimate for S_r, 100 sets of random points are used for each image, each generating an S_r value. The median S_{rm} of this set of S_r values for each image is used for further analysis. Ratios S_h/S_{max} and S_{rm}/S_{max} and the medians for each of these are calculated. The ratio of these medians is then calculated. Higher ratios mean that saliency values around fixation points are greater than saliency values around random end points, showing that the model can predict human gaze locations in the image better than expected by chance.

Image shuffling. Shuffling is a control analysis where instead of using gaze points for the image in consideration, gaze points of another randomly chosen image are used. The ratio of medians analysis stated above is done using the saliency maps from one image and gaze data from the randomly chosen image. These results are then compared to the results when using the saliency maps and gaze data for the same image.

Differences between S_h and S_{rm} were evaluated using a statistical sign test with a significance level of 0.0001 for both the full and new models for the cases with and without shuffling. Also, the same statistical test was carried out between the S_h values with and without shuffling to see if the S_h values with shuffling are significantly less than the S_h values without shuffling.

2.3.2. Analysis using normalized scanpath salience (NSS) [26]. This method normalizes the salience map to have a zero mean and unit standard deviation. For each point corresponding to the fixation locations, the normalized salience value is extracted and the mean of all these extracted values is calculated. This mean is the normalized scanpath salience (NSS) value. If the NSS value is greater than zero, there is a greater correspondence between the salience maps and gaze fixation points than expected by chance. The NSS value of zero would mean there is no such correspondence and a value of less than zero would mean there is anti-correspondence between the salience maps and human fixations. To verify this in practice, chance values are calculated by creating a map with a uniform distribution at the same resolution as the saliency map instead of the actual salience map and calculating NSS in the same manner as stated above. We calculated the NSS values for all gaze data points by taking a region of diameter 5.6° around each fixation point in order to avoid any fixation drifts and minor calibration offset effects. Here again, random map generation was carried out 100 times for each image.

For both the full and new models, NSS values obtained using salience maps were compared to the NSS values obtained using random maps (paired *t*-test with a significance level of 0.0001).

3. Results

3.1. Saliency maps

Figure 2 shows the salience maps generated by the new and full models for the same input image. Figure 2(a) shows an example of an input image, the saturation, intensity and edge conspicuity maps and the final salience map created by the new model. Figure 2(b) shows the salience map computed by the full model for the same input image. On comparing the final saliency maps from the new and the full models, we can observe that the salient image areas (e.g. the curb, the plant, etc) are similar in the outputs of both models. The new model works with a lower resolution image (320 × 240 pixels) than the full model (640 × 480 pixels) resulting in coarser maps when compared to the full model.

Saturation and edge conspicuity maps (figure 2(a)) enhance objects with more saturated hues and darker colors and objects with prominent edges respectively whereas the intensity conspicuity map enhances objects with intensity contrast in the image frame. The process of creating the feature maps and normalizing them can lead to certain pixels not being enhanced in the final conspicuity map.

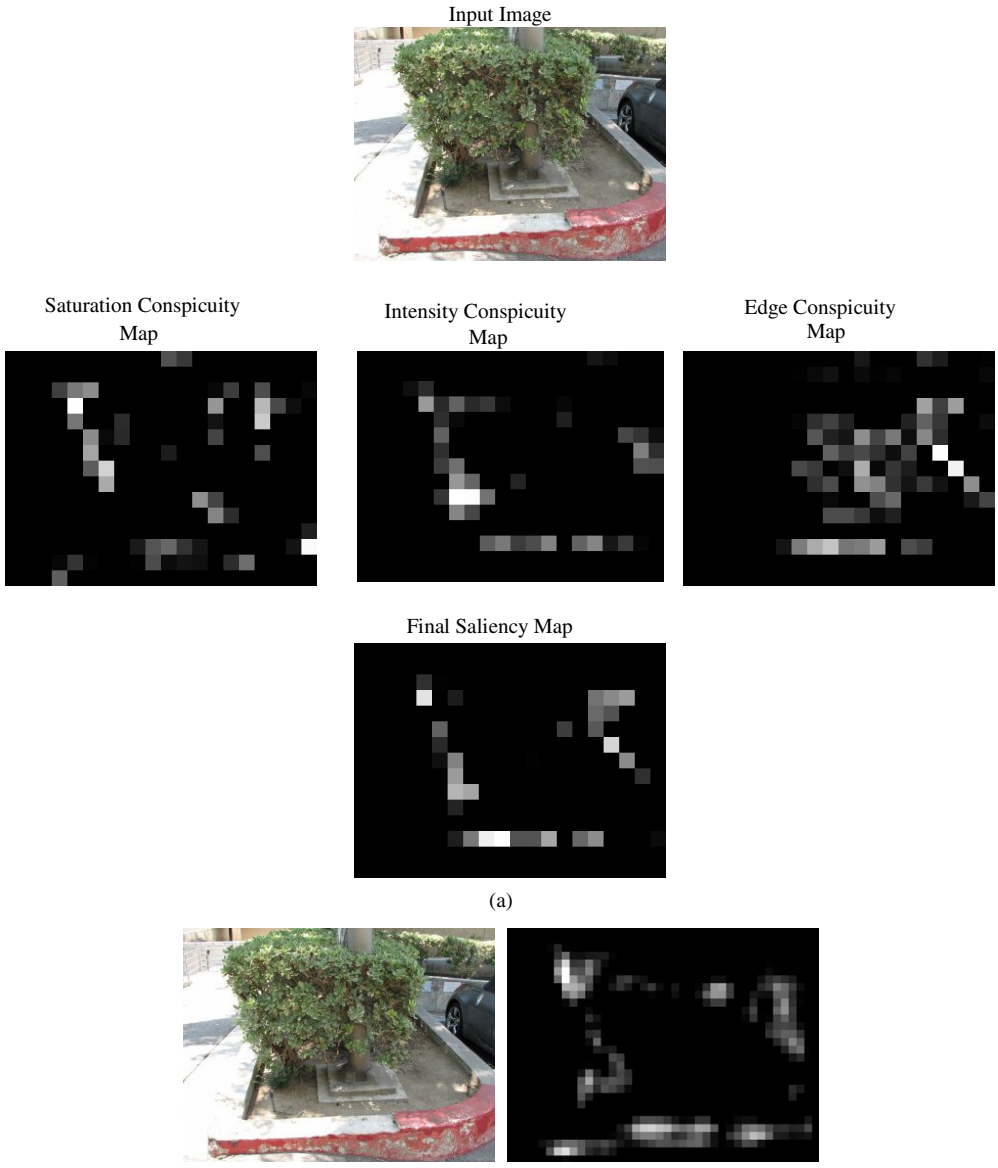

Figure 2. Salience maps created by the new model and the full model for the same input image: (a) conspicuity maps for saturation, intensity and edge along with the final salience map created by the new model for an example input image; (b) final salience map created by the full model for an example input image.

3.2. DSP implementation results

The various modules of each model and the time required to process one frame are stated in table 1. As stated earlier, for the full model, only the intensity stream which is one of seven different streams has been implemented on the DSP.

Execution time results from table 1 show that a single image frame takes 0.84 s to be processed by the new model whereas 14% of the old model (only the intensity stream) takes 1.53 s to process the same image. This shows that the implementation of the new model is computationally more efficient. The estimated time for the full model to execute one frame can be calculated by multiplying the time for the intensity stream execution by a factor of 7. This is because the intensity stream is one of seven similar streams in terms of the computational complexity in the original saliency algorithm. This implies that the implementation of the new model is approximately ten times faster than the implementation of the full model.

Optimization can lead to better results as can improvements in processor speed and power consumption. However, in wearable computing systems, increased algorithm efficiency will always translate into lower power consumption. Even the unoptimized implementation of the new model can execute in less than 1 s, which is a reasonable response time to a user request for information.

3.3. Model validation using gaze data

3.3.1. Analysis with the ratio of medians method [25]. Figure 3 shows two examples of input images with salience maps from the new model and the full saliency model. Points in images

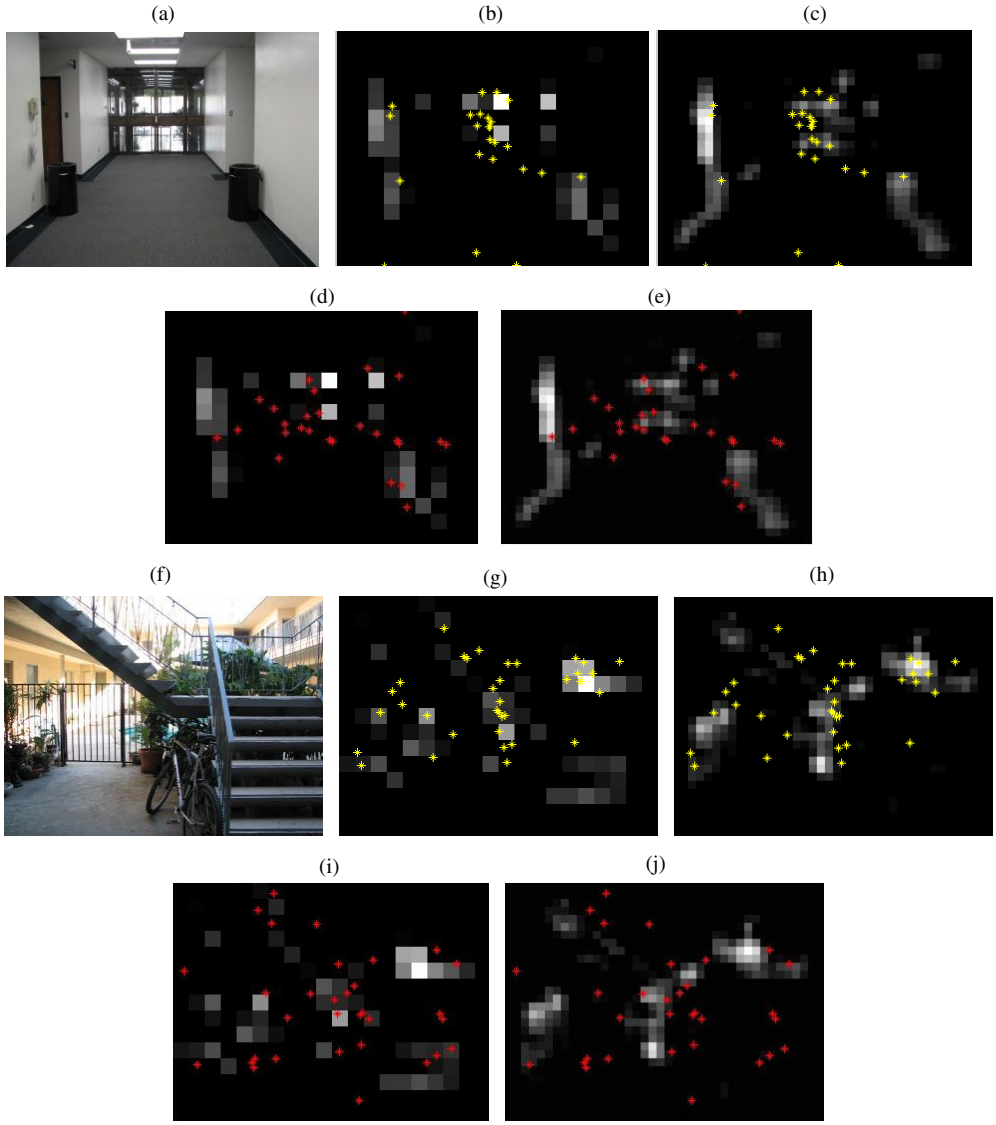

Figure 3. Input image (a and f), salience maps from new model (b and g) and full saliency model (c and h) with the dots depicting gaze fixation points, and salience maps from the new model (d and i) and full saliency model (e and j) with dots depicting data points after shuffling.

Table 1. Time in seconds for computation of different modules in the new model and intensity stream of the full model on the TMS320 DM642 DSP.

Functions	New model implementation (time in seconds)	Intensity stream from the full model implementation (time in seconds)
YCbCr → RGB	0.1250	–
RGB → HSI	0.0320	–
Gaussian pyramids (intensity and saturation for new model)	0.0647	0.0695
Laplacian pyramids	0.0303	–
Center–surround maps	0.0027	0.0158
Normalization function	–	0.6062
(at different scales)	–	0.0757
	0.0096	0.0113
Entire algorithm (s)	0.8416	1.5373
Frames/second	1.1882	0.6505

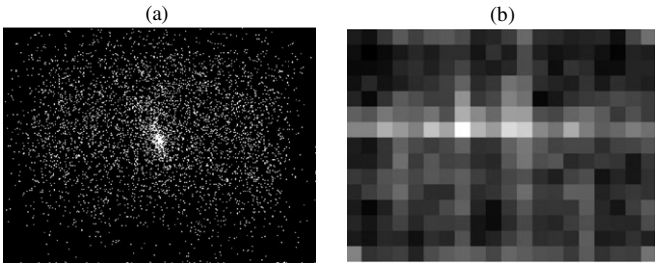

Figure 4. Gaze distribution of all subjects over all images (a) and the average salience map from the salience maps of all images (b) in the entire data set.

Table 2. Data analysis of human gaze data with salience maps created by the new model and the full saliency model using the ratio of medians method.

	Median (S_h/S_{max})	Median (S_{rm}/S_{max})	Ratio of medians	Sign test (S_h and S_{rm})
New model	0.3647	0.1020	3.5769	$p < 0.0001$
Full model	0.4352	0.2457	1.7714	$p < 0.0001$
Image shuffling				
New model	0.2275	0.1059	2.1481	$p < 0.0001$
Full model	0.3256	0.2511	1.2970	$p < 0.0001$

(b), (c), (g) and (h) depict gaze fixation points from human subject data and the points in images (d), (e), (i) and (j) depict data points obtained by shuffling (gaze points from another image).

Table 2 shows analysis of the full and the new models for the actual gaze data and randomly distributed gaze points. Both the full and the new models have ratios which are significantly above chance (sign test with a significance level of 0.0001 carried out between the S_h and S_{rm}, chance = 1) indicating that both models predict better than chance where human observers will look. The ratio of medians calculated by the new model is higher than the full model which shows that the new model outperforms the full model in this case. The maps from the full model are slightly denser than the maps from the new model as seen in the comparison between figures 3(i) and (j). This results in the overall median values of the full model being greater than the new model.

The shuffled image analysis results are also shown in table 2. A statistical sign test with a significance level of 0.0001 between the S_h values with and without shuffling indicates that the S_h values without shuffling are significantly higher than the S_h values with shuffling. The shuffled analysis shows that the median values and ratios are lower than when the saliency maps and gaze data correspond, but the ratios are statistically greater than one, that is, better than chance. This discrepancy can be explained by the center-bias effect present in the average salience map of all images as well as in the gaze data of subjects. When looking at unfamiliar images, subjects often start looking at the center and then proceed to examine the peripheral areas. Subjects are asked to look at the center of a test image after every three images for calibration purposes as mentioned before which could also add to their initial fixation being centrally biased. Finally, due to potential photographer bias (having interesting objects in the center of the image), the average of all the salience maps in the input image data set also has a center bias. Figure 4 shows the center-bias in the gaze data as well as the average salience map.

Table 3. Data analysis of human gaze data with salience maps created by the new model and the full saliency model using the ratio of medians method for the image data set after removing images with a center bias in the gaze data and/or the salience maps.

	Median (S_h/S_{max})	Median (S_{rm}/S_{max})	Ratio of medians	Sign test (S_h and S_{rm})
New model	0.3490	0.1020	3.4231	$p < 0.0001$
Full model	0.4278	0.2519	1.6985	$p < 0.0001$

To investigate the finding that saliency and gaze were correlated even with image shuffling, a center-bias analysis of the gaze points from subjects as well as the average salience map was done. Based on Tatler's analysis [27], for each image, the number of gaze points falling into the central 15° was counted and compared to the number of gaze points falling into the rest of the image areas which are referred to as peripheral areas. If the number of gaze points in the central region was greater than the number in the peripheral regions, there was a center bias in gaze data. Calculations show that 26% of the images used in our study have a center bias in the subject gaze data. Similarly, the number of pixels whose grayscale level is the maximum value of the average salience map is calculated in the central and peripheral areas of the average salience map. If the number of such maximum grayscale valued pixels is greater in the center than in the periphery, there is said to be a center bias. Figure 4(b) shows the average saliency data from all images, indicating central bias. The bias in the subject gaze while viewing these unseen natural images and the bias in the salience maps due to the photographer bias may be a reason behind the ratio of medians being greater than 1 even with image shuffling. In general, if the combination of the shuffled gaze data set and the salience map is such that both have an overlap, the ratio for such combinations will be greater than 1.

Analysis was repeated after removing images with a central bias. Table 3 shows these results. The median values for S_h/S_{max} and S_{rm}/S_{max} are very close to those obtained with the entire set of images including the ones with a center bias. As before, a sign test with a significance level of 0.0001 carried out between the S_h and S_{rm} shows that S_h values are significantly higher than S_{rm} values.

3.3.2. Analysis using normalized scanpath salience (NSS) [26]. Figure 5 shows two examples of an input image with salience maps from the new model and the full saliency model. The figure also shows a random map created from a uniform distribution at the same resolution as the salience maps for the new and full models. The dots in the salience and random maps represent the gaze fixation points of human observers.

The NSS results are shown in table 4 for the new model as well as the full model. NSS values for both models with salience maps are greater than zero whereas the NSS value for the random model is close to zero. The results for the analysis on the data set of images after removing images with a gaze or salience map center bias are shown in table 5.

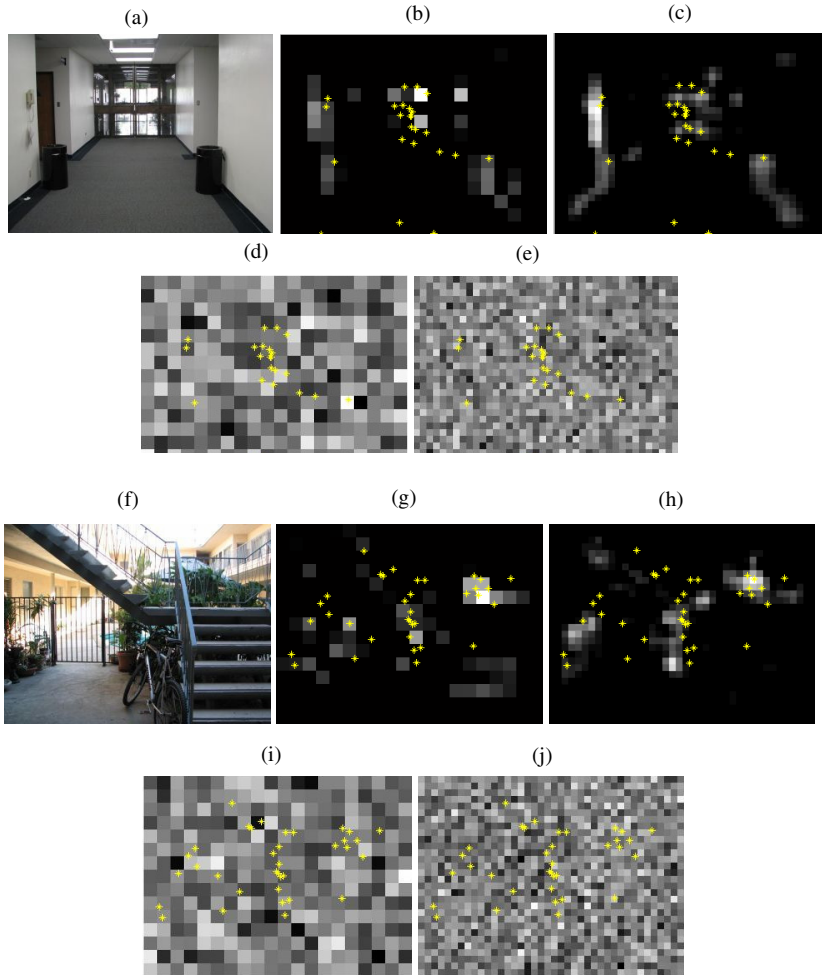

Figure 5. Input image (a and f), salience maps from the new model (b and g) and the full saliency model (c and h); uniform distribution random map with the dots depicting the gaze fixation points of the human subjects for the new model (d and i) and for the full model (e and j).

Table 4. Data analysis of human gaze data with salience maps created by the new model and the full model using the normalized scanpath salience method.

	For salience map NSS ± SEM	For random map NSS ± SEM	Paired *t*-test
New model	0.4310 ± 0.0113	$-4.813 \times 10^{(-4)} \pm 0.0005$	$p < 0.0001$
Full model	0.4758 ± 0.0098	$-5.077 \times 10^{(-4)} \pm 0.0005$	$p < 0.0001$

Table 5. Data analysis of human gaze data with salience maps created by the new model and the full model using the normalized scanpath salience method for the image data set after removing images with a center bias in the gaze data and/or the salience maps.

	For salience map NSS ± SEM	For random map NSS ± SEM	Paired *t*-test
New model	0.4153 ± 0.0104	$1.2311 \times 10^{(-4)} \pm 0.0006$	$p < 0.0001$
Full model	0.4746 ± 0.0093	$1.9836 \times 10^{(-4)} \pm 0.0005$	$p < 0.0001$

For both cases, a paired *t*-test with a significance level of 0.0001 shows that the NSS values obtained using salience maps are significantly different than the NSS values obtained using random maps, meaning there is greater correspondence between salient regions detected by the salience maps and human fixations than expected by chance.

4. Discussion

We present a computationally efficient model of bottom-up saliency detection based upon an earlier saliency model [11–15]. Good correspondence is noted when comparing regions that the algorithm predicts as salient to regions gazed

at by human subjects when looking at a set of images. Also comparing the salient regions to random gaze points shows that the model predicts salient regions at a rate better than what would be expected by chance. We have validated our algorithm with sighted observers viewing images on a computer screen while seated. The subjects are shown unfamiliar scenes to limit unbiased gaze patterns. Using a set of 150 images and gaze data from five subjects, results show comparable performance between the new and the full models. An unoptimized implementation on the TMS320 DM642 DSP shows that the proposed model can process at a rate of 1 frame per second which is approximately ten times faster than the full model. Since this algorithm is eventually hoped to be run on a wearable computing platform, efficiency is a critical factor.

The model is proposed as the core of an image processing algorithm designed to provide visual prosthesis patients with information about the areas outside the visual field of the implant. Such an algorithm could be utilized in a number of ways. During navigation and ambulation, the user might want to know about obstacles or signs (for example, an exit sign). Other times, the user might be searching for an object of interest. While it is possible to design specific algorithms tailored to each task, there are advantages to a bottom-up approach. Unlike top-down algorithms that require *a priori* information, bottom-up algorithms do not require any training. Also, a bottom-up algorithm may allow the user to identify objects and understand surroundings using remnant vision and contextual cues. Nevertheless, it is possible that a top-down algorithm may be needed for specific tasks such as objective recognition, particularly where vision is very poor. Frintrop *et al* proposed a saliency implementation based on ten information streams and a five level image pyramid scheme for robotics [20]. Walther *et al* proposed a bottom-up implementation based on the full model with an added feedback module to detect the extent of an attended object [28]. Both the groups combined their bottom-up implementations by using top-down information based on feature detection and/or object recognition with the salient regions [29, 30].

To be effective for a retinal prosthesis implant patient, more work is required to understand what functions are important for these patients. Training will also be required to best utilize information provided by the algorithm. Also, it is unclear if patients can learn to take advantage of the additional information or if they will prefer to receive unfiltered video data and make their own judgments about object importance. Task-dependent processing may be the best approach. For obstacle avoidance and route planning, visually impaired people are likely to be more interested in large objects obstructing their path versus small details of the environment around them. In such a case, a saliency algorithm may be adequate. For an object detection task, smaller details may be important discriminating clues to aid successful task completion and a top-down object recognition algorithm may be required. Any additional algorithm will require more computing power so additional benefit will come at a cost. The extra computing load can be limited by applying object recognition only in the small region identified as salient. This would also eliminate the need for the entire image frame to be processed in smaller parts by the object recognition algorithm to find the various objects.

In summary, a computationally efficient image processing algorithm has been analyzed and identifies parts of an image that human observers also deem salient. This algorithm has the potential to enhance low vision, particularly when visual field is restricted. When used with a retinal prosthesis, the algorithm can be implemented on the retinal prosthesis' existing camera and wearable computing platform. Critical questions remaining to be answered by human testing include how quickly people learn to utilize the algorithm and the benefit provided.

Acknowledgments

This research was performed at the Biomimetic Microelectronics Systems Engineering Research Center. This material is based on the work supported by the National Science Foundation under grant no EEC-0310723. This work was also supported (LI) by grants from the National Science Foundation, the Defense Advanced Projects Agency and the Army Research Office. The views, opinions and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Agency or the Department of Defense. We would also like to thank Dr Armand Tanguay and his student Benjamin McIntosh for the Arrington Research equipment to record gaze data and David Berg at iLab for his help with the saccade and fixation filtering software.

References

[1] Margalit E and Sadda S R 2003 Retinal and optic nerve diseases *Artif. Organs* **27** 963–74
[2] Kolb H, Fernandez E and Nelson R 2002 *Webvision: The Organization of the Retina and the Visual System*
[3] de Balthasar C *et al* 2008 Factors affecting perceptual thresholds in epiretinal prostheses *Invest. Ophthalmol. Vis. Sci.* **49** 2303–14
[4] Yanai D *et al* 2007 Visual performance using a retinal prosthesis in three subjects with retinitis pigmentosa *Am. J. Ophthalmol.* **143** 820–7
[5] Gekeler F, Sachs H, Szurman P, Guelicher D, Wilke R, Reinert S, Zrenner E, Bartz-Schmidt K and Besch D 2008 Surgical procedure for subretinal implants with external connections: the extra-ocular surgery in eight patients *ARVO 2008 Annual Meeting* ARVO Abstract-4049
[6] Humayun M S 2009 Preliminary results from Argus II Feasibility study: s 60 electrode epiretinal prosthesis *ARVO 2009 Annual Meeting* ARVO Abstract 4744
[7] Dagnelie G 2008 Psychophysical evaluation for visual prosthesis *Ann. Rev. Biomed. Eng.* **10** 339–68
[8] Asher A *et al* 2007 Image processing for a high-resolution optoelectronic retinal prosthesis *IEEE Trans. Biomed. Eng.* **54** 993–1004
[9] Hallum L E, Cloherty S L and Lovell N H 2008 Image analysis for microelectronic retinal prosthesis *IEEE Trans. Biomed. Eng.* **55** 344–6

[10] Boyle J R, Maeder A J and Boles W W 2008 Region-of-interest processing for electronic visual prosthesis *J. Electron. Imaging* **17** 013002
[11] Itti L, Koch C and Niebur E 1998 A model of saliency-based visual attention for rapid scene analysis *IEEE Trans. Patt. Anal. Mach. Intell.* **20** 6
[12] Itti L 2000 Models of bottom-up and top-down visual attention *PhD Thesis* California Institute of Technology, Pasadena p 216
[13] Itti L and Koch C 2000 A saliency-based search mechanism for overt and covert shifts of visual attention *Vis. Res.* **40** 1489–506
[14] Itti L and Koch C 2001 Computational modelling of visual attention *Nat. Rev. Neurosci.* **2** 194–203
[15] Itti L and Koch C 2001 Feature combination strategies for saliency-based visual attention systems *J. Electron. Imaging* **10** 161–9
[16] Kandel R, Schwartz J and Jessel T 2000 *Principles of Neural Science* 4th edn (New York: McGraw-Hill)
[17] Wolfe J M, Cave K R and Franzel S L 1989 Guided search: an alternative to the feature integration model for visual search *J. Exp. Psychol. Hum. Percept. Perform.* **15** 419–33
[18] Wolfe J M 1994 Guided search 2.0: a revised model of visual search *Psychon. Bull. Rev.* **1** 202–38
[19] Koch C and Ullman S 1985 Shifts in selective visual attention: towards the underlying neural circuitry *Hum. Neurobiol.* **4** 219–27
[20] Frintrop S 2005 *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search* (Germany: University of Bonn)
[21] Greenspan H, Belongie S, Goodman R, Perona P, Rakshit S and Anderson C H 1994 Overcomplete steerable pyramid filters and rotation invariance *IEEE Computer Vision and Pattern Recognition (Seattle, Washington)*
[22] Burt P J and Adelson E H 1983 The Laplacian pyramid as a compact image code *IEEE Trans. Commun.* **31** 532–40
[23] Adelson E H, Anderson C H, Bergen J R, Burt P J and Ogden J M 1984 Pyramid methods in image processing *RCA Engineer* **29** pp 33–41
[24] Texas Instruments, *TMS320DM642 Video/Imaging Fixed-Point Digital Signal Processor (Rev. L)*
[25] Itti L 2005 Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes *Vis. Cogn.* **12** 1093–123
[26] Peters R J, Iyer A, Itti L and Koch C 2005 Components of bottom-up gaze allocation in natural images *Vis. Res* **45** 2397–416
[27] Tatler B W 2007 The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions *J. Vis.* **7** (4) 1–17
[28] Walther D and Koch C 2006 Modeling attention to salient proto-objects *Neural Netw.* **19** 1395–407
[29] Frintrop S, Backer G and Rome E 2005 Goal-directed search with a top-down modulated computational attention system *Proc. Ann. Meeting of the German Association for Pattern Recognition (DAGM'05) (Wien, Austria)*
[30] Walther D 2006 Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics *PhD Thesis* California Institute of Technology