Optical Engineering 44(10), 107004 (October 2005)

Vision-model-based image foveation and motion estimation

Yitzhak Yitzhaky, MEMBER SPIE
Ben-Gurion University
Department of Electro-Optics Engineering
P.O. Box 653
Beer-Sheva 84105, Israel

Eli Peli, MEMBER SPIE
The Schepens Eye Research Institute
20 Staniford Street
Boston, Massachusetts 02114-2500
E-mail: eli@vision.eri.harvard.edu

Abstract. Foveated imaging systems applicable in various single-user displays mimic the visual system's image structure, where resolution decreases gradually away from the fovea. The main benefit is the low average image resolution while maintaining high resolution at the center of the gaze. When the end user is a human observer, it is advantageous for the foveation process to closely match the visual system parameters. This work directly applies a multichannel model of the visual system to form foveated images. A systems-engineering approach applied to the vision model produces quantitative image spectral content across the visual channels. Foveated images are constructed according to the contrast threshold and image content calculated at different eccentricities. Also, variable-resolution feature detection (edge and bar) that corresponds to early visual processing is produced, based on the available image content across the channels. Motion between shifted foveated images (required in applications such as image compression and motion compensation) is estimated using either the foveated images or the detected feature images. Results using several similarity metrics and imaging conditions show that reliable motion estimation can be achieved, while features with nonsimilar resolutions (different scales) are matched. *© 2005 Society of Photo-Optical Instrumentation Engineers.*
[DOI: 10.1117/1.2084667]

Subject terms: foveated imaging; multiresolution; motion estimation; vision model; image registration; head-mounted displays.

Paper 040702R received Sep. 24, 2004; revised manuscript received Mar. 27, 2005; accepted for publication Mar. 29, 2005; published online Oct. 20, 2005.

1 Introduction

The spatial resolution of the human visual system is highest at the center of the retina (fovea), and decreases rapidly away from it. Formation of images that resemble this spatially variant property is called image foveation.[1,2] In space-variant foveating systems, the area of interest (AOI) in the image is maintained at high resolution, while areas away from it, where precise detail may be less critical, are coded at lower resolution. With this image structure, the average resolution is considerably lower than in the standard uniformly sampled high-resolution image structure. Thus, considerable increase in compression ratio can be achieved.[1,2] The low average resolution of foveated images, while maintaining high resolution in the AOI, motivated research aimed at increasing the efficiency of data processing for transmission and visualization. Several foveation models and techniques have been proposed.[3–7] Applications of foveated images include robotic active-vision systems,[8] video conferencing,[9] and driving or flight simulations.[10] Foveated imaging is most effective when the direction of the gaze of the observer is tracked, so that the highest resolution region at the display can be kept aligned with the user's fovea.[10–12]

The human vision system uses sequences of shifted retinal images of the scene to derive sharp high-resolution views of multiple areas of an image. These shifts are attained through a series of fast saccadic eye movements between AOIs. Similarly, shifted images may be produced by imaging systems such as head-mounted displays with head-mounted cameras or with remote robot sensors. A significant step for reducing frame-to-frame redundancies in compression is an estimation of the extent of the movement between the images or subimages (usually termed motion estimation).[13] Motion estimation in a foveated system has to be performed from images with different foveal locations. Variable resolution may complicate or impede motion estimation in such systems.

In the visual process, it was suggested that the foveated retinal image information is employed together with extraretinal information (neural control signals and eye muscle mechanical sensors' signals) to achieve perception stability in spite of saccades.[14–16] In this scheme, the image information is used for the final fine registration of pre- and postsaccadic images.

Object features (contour edges or bars) are believed to be extracted early in the visual process.[17,18] Edges are image features associated with the transition from dark to bright luminance across the feature. Bars are thin bright or dark features on a contrasting background.[19] Such features are also widely utilized in various computer applications, because they provide important information about object boundaries. They are invariant to luminance changes and give accurate information about the spatial location of objects, while occupying only small portions of the image space. (The relatively homogeneous areas of the image ob-

0091-3286/2005/$22.00 © 2005 SPIE

jects or background occupy most of the space). Image registration using edge features was shown to be faster in certain techniques,[20] and less affected by different gray-level characteristics of the matched images.[21] Therefore, it may be a more effective approach for motion estimation, where computing time or resources are limited. In foveated images, however, the low resolution at the higher eccentricities may decrease the spatial accuracy of such features.

In this work, we address a few aspects of foveation and its applications in imaging systems. We use a systems-engineering approach[22] to analyze a multichannel foveation vision model,[23–27] where a uniform high-resolution scene is the system's input and the foveated image is the output. Image energy is quantitatively evaluated with regard to the visual channels (spatial scales) and the angular distance from the fovea (retinal eccentricity). The quantitative results are used to implement feature extraction from space-variant foveated images.

To examine the use of edge features in a foveated system environment, we extend a vision-model-based feature detection method[28] to include the spatially variant resolution properties of the visual system. Variable-resolution feature detection results are then used to estimate the motion between shifted foveated images. In this process, features obtained at different levels of resolution are registered.

The rest of the work is organized as follows. Section 2 presents a systems-engineering analysis of the multichannel vision model. In the first subsection, the vision model is described, and in the second, the model is analyzed, providing a quantitative measure of the channels' transfer of image energies at different eccentricities. Vision-model-based image foveation is presented in Sec. 3. In Sec. 4 a vision-model-based feature detection method is extended to include different eccentricities with different resolution levels (rather than a single resolution for the whole image). Section 4.1 summarizes the original feature detection method (with space-invariant resolution), and Sec. 4.2 presents the space-variant extension. Motion estimation from foveated images, using image registration with several similarity metrics, is presented in Sec. 5. Results of image foveation, foveated feature detection, and motion estimation are shown in Sec. 6. Summary and conclusions are presented in Sec. 7.

2 Systems-Engineering Analysis of Multichannel Vision Model

In systems engineering, the output signal of a system is related to its input signal by the system property, usually referred to as its *transfer function*. In this section, the multichannel vision model is described briefly, and then it is analyzed with regard to the input image energy distribution across the channels and the system model properties. The available image energies calculated at different channels and eccentricities are used later in the feature extraction process (Sec. 4.2).

2.1 *Multichannel Vision Model*

Empirical research of the visual system suggests a multichannel vision model, with the channel properties varying as a function of their spatial location (mainly eccentricity) with respect to the fovea,[23,24] and in which several spatial frequencies and orientationally tuned channels exist at each

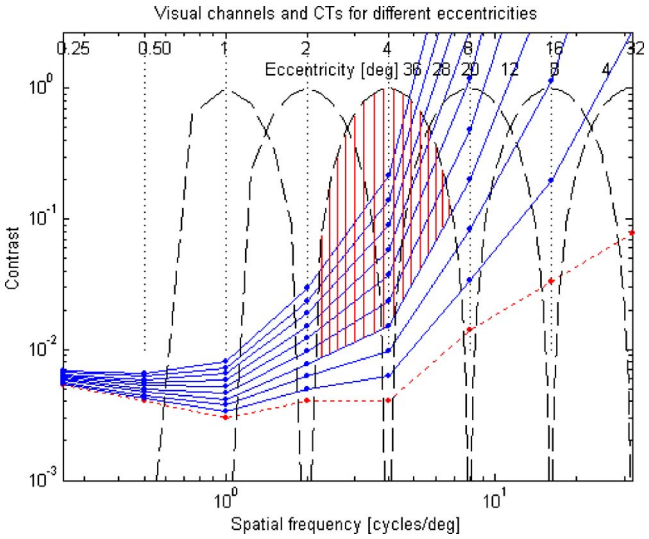


Fig. 1 Cross sectional profiles of CTs $Th(\theta, r)$ at different eccentricities, with the maximum possible observed contrasts $MC_i(r)$ that correspond to the multichannel filters, MTFs $C_i(r)$ of the visual model. The lower dotted line is the CT at the fovea, the continuous lines are CTs at eccentricities from 4 deg (the second lowest) to 36 deg (the upper), and the dashed lines are the receptive channels' responses according to Eq. (1). The area marked by the vertical lines is the $CMTFA_{12,2}$ (at eccentricity 12 deg and the channel with 4 cycles/deg center frequency).

position in the visual field.[25] The filters are about one octave wide in the spatial frequency domain, separated by one octave in their center frequency, and are also orientation selective.[29] The model used here was applied only to spatial frequency and not to orientation; however, it could be extended easily to orientation.[28]

The i'th order channel (bandpass filter) C_i (Fig. 1) applied in the frequency domain can be approximated in an engineering model of the system as[30]:

$$C_i(r) = \begin{cases} 0.5[1+\cos(\pi \log_2 r - \pi i)], & 2^{i-1} \leqslant r \leqslant 2^{i+1} \\ 0, & \text{elsewhere} \end{cases}, \tag{1}$$

where r is the radial spatial frequency. Other one-octave-wide filters such as the log Gaussian[31] may serve equally well.

Peli, Yang, and Goldstein[32] developed a mathematical model for the variation of the contrast threshold (CT)—the reciprocal of the contrast sensitivity function—as a function of spatial frequency and eccentricity from the fovea. Two properties of the visual system's response to changes in frequency and eccentricity specify the model. The first is the exponential rise with eccentricity of measured CTs at any one spatial frequency. The second, called *contrast constancy*,[33] accounts for the invariance of the appearance of (suprathreshold) objects with changes in the size of their retinal images, which may result from changes in their distance from the eye. The inverse relation between retinal size and spatial frequency, while contrast constancy is maintained, suggests that CTs vary as a product of the spatial frequency and the eccentricity. The CT, Th, at eccen-

Table 1 Values of $CMTFA_{\theta,i}$ for channels and eccentricities shown in Fig. 1. All the values are multiplied by 10^3.

Center frequency [cyc/deg]	0.5	1	2	4	8	16	32
Eccentricity [deg]							
0	0.3012	1.217	4.87	19.3	75	284	526
2	0.3011	1.216	4.86	19.2	72	232	294
4	0.3010	1.215	4.85	19.0	67	148	0
8	0.3008	1.212	4.83	18.4	51	13	0
12	0.3005	1.210	4.79	17.3	31	0	0
16	0.3002	1.206	4.75	15.8	15	0	0
20	0.2998	1.202	4.69	13.6	3	0	0
24	0.2994	1.197	4.60	11.1	0	0	0
28	0.2989	1.191	4.49	8.7	0	0	0
32	0.2984	1.183	4.34	6.6	0	0	0
36	0.2978	1.174	4.15	5.0	0	0	0

tricity θ, and (radial) spatial frequency r is expressed by this model as

$$\ln[Th(\theta,r)] = a\theta r + \ln[Th(0,r)], \tag{2}$$

where a is a constant and $Th(0,r)$ is the foveal threshold at spatial frequency r. This model was successfully fitted to CT measurements at different frequencies as a function of eccentricity,[32] reported in various studies,[34–36] and was recently verified empirically.[37] This model can be applied to derive the cortical image representation obtained from different retinal eccentricities, given the foveal CT and the eccentricity constant a.

2.2 *System Analysis*

The modulation transfer function (MTF) of an imaging system describes the system's amplitude response to sinusoidal inputs at different spatial frequencies. The MTF ranges between 1 (at spatial frequencies where signal transfer is maximum) and 0 (at frequencies with no signal transfer). The perceptual quality of an image transferred through an imaging system is limited by both the MTF of the system and the CT of the observer's visual system.[22,38] The area enclosed by the graphs of the MTF and the CT (termed MTFA)[22] specifies the space of all contrast values at all spatial frequencies that can be observed due to the combination (limitations) of the imaging system (MTF curve) and the observer (CT curve). To obtain the actual observed image, the spectral distribution of the input image is multiplied by the system's MTF, and the result is thresholded by the CT. Accordingly, by considering the CTs at different eccentricities in the multichannel vision model, we can quantitatively calculate the possible signal contrast range that can be received by the observer through every channel i at a given eccentricity θ, according to the area captured between the filter and the CT. Hence, we define a single channel MTFA (CMTFA) as:

$$\mathrm{CMTFA}_{\theta,i} = \int_{2^{i-1}}^{2^{i+1}} [MC_i(r) - Th(\theta,r)]dr, \quad C_i(r) > Th(\theta,r), \tag{3}$$

where $MC_i(r)$ is the maximum possible observed contrast at frequency r, obtained by multiplying the channel's MTF $C_i(r)$ (that ranges between 0 to 1) by an input contrast that equals 1 at all frequencies. Figure 1 is a 1-D graphical representation of the intersection of the CTs at different eccentricities with the multichannel visual model. The lower dotted line is the CT at the fovea, the continuous lines are CTs at increasing eccentricities from 4 deg (the second lower) to 36 deg (the upper) computed with Eq. (2), and the dashed lines are the receptive channels (filters) according to Eq. (1). The $\mathrm{CMTFA}_{\theta,i}$ is expressed in this figure as the area captured between the maximum possible observed contrast $MC_i(r)$ with center frequency $r=2^i$ and the CT of eccentricity θ. For example, $\mathrm{CMTFA}_{12,2}$ is marked in Fig. 1 by vertical lines. Table 1 shows values of $\mathrm{CMTFA}_{\theta,i}$ for the eccentricities and channel orders shown in Fig. 1.

The CMTFA values in Table 1 reflect known properties of the visual system. The CMTFA should increase with the channel order, since the (linear) bandwidths of the channels increase logarithmically as the order increases. However, the increase of the CMTFA with the order is smaller than the logarithmic increase of the filter size, since the CT also increases with the channel order. These CMTFA values can be explained by the properties of the visual system that

Table 2 A typical example of image energies that pass through each channel at different eccentricities. All values are multiplied by 10^3. The bold numeral cells show typical sets of channels $\Re_\theta$, used in the feature-extraction process for different eccentricities. Note that four channels are used for low eccentricities and only three channels for the higher eccentricities.

Center frequency [cyc/deg]	0.5	1	2	4	8	16	32
Eccentricity [deg]							
0	213.0	167.7	**106.0**	**98.9**	**79.9**	**49.9**	1.2
2	213.0	167.7	**106.0**	**98.2**	**71.6**	**16.4**	0.0
4	213.0	167.7	**106.0**	**97.2**	**57.1**	0.2	0.0
8	213.0	167.7	**105.9**	**93.2**	**16.5**	0.0	0.0
12	213.0	**167.7**	**105.8**	**84.2**	0.2	0.0	0.0
16	213.0	**167.6**	**105.6**	**67.4**	0.0	0.0	0.0
20	213.0	**167.6**	**105.3**	**42.6**	0.0	0.0	0.0
24	213.0	**167.6**	**104.8**	**21.2**	0.0	0.0	0.0
28	213.0	**167.5**	**104.1**	**6.9**	0.0	0.0	0.0
32	**213.0**	**167.5**	**103.0**	0.4	0.0	0.0	0.0
36	**213.0**	**167.4**	**101.3**	0.0	0.0	0.0	0.0

evolved in response to spatial characteristics of natural scenes.[32] Though the important and useful information in the image is mainly represented by the boundaries of the objects, the areas of the boundaries are obviously much smaller than the nearly homogeneous parts in the image. The spectra of real (natural) images are known to rapidly decrease with regard to the spatial frequency,[39] usually proportional to $1/r$. Thus, the boundaries are represented with a much lower energy. Therefore, the visual system enables wider receiving bandwidths for the higher frequency channels that process the boundary (high frequency) information. This is particularly true at the lower eccentricities, mainly in the fovea. At higher eccentricities (used mainly for navigation and danger detection), a rough knowledge of objects' existence (available by their lower frequencies or just by movement) is sufficient, and therefore, no higher frequencies image-contrast receiving capability (represented by the CMTFA values) is available there. This organization of the visual system is necessary to efficiently manage the large amount of optical information constantly reaching it.

The interaction of the vision model with a sample input image is demonstrated quantitatively in Table 2. The table presents an example of the typical image energies that would be transmitted through each channel at different eccentricities. In this table, the image section [shown in Fig. 2(a)] was filtered by the visual channels of the model at different eccentricities. It can be seen from Table 2 that although the CMTFAs of the higher order channels are bigger (Table 1), image energies that get through these channels are smaller as a result of the diminishing input image energy at higher spatial frequency.

3 Vision-Model-Based Image Foveation

In the image foveation process, the observed scene is transferred (bandpass filtered) through the channels, forming images at spatial scales that can be indexed according to the order of the channel i. Each scale i, with spatial frequency pass band ranging from 2^{i-1} to 2^{i+1}, is thresholded by the CT, $Th_{\theta,i}$ [obtained from Eq. (2)], approximated by the value of the CT measured at the spatial frequency $r=2^i$:

$$Th_{\theta,i} = \exp[a\theta 2^i + \ln(Th_{0,i})], \tag{4}$$

where θ is the eccentricity relative to the foveal location. The foveated image is constructed by summing all the suprathresholded scales. (The threshold contrast's index is converted from the cyc/deg units with which the threshold is measured, to cyc/image units, using the image's angular span in degrees.)

4 Extension of a Vision-Model-Based Feature Detection Method to Different Eccentricities

Feature (edge and bar) detection is considered to be a basic low-level vision task. A formulation of feature detection in a foveated system environment is presented here. For this purpose, a vision-model-based feature detection technique[28] is extended and applied to different eccentricities. First, a brief summary of the basic underlying technique is presented, followed by the extension to spatially variant system.

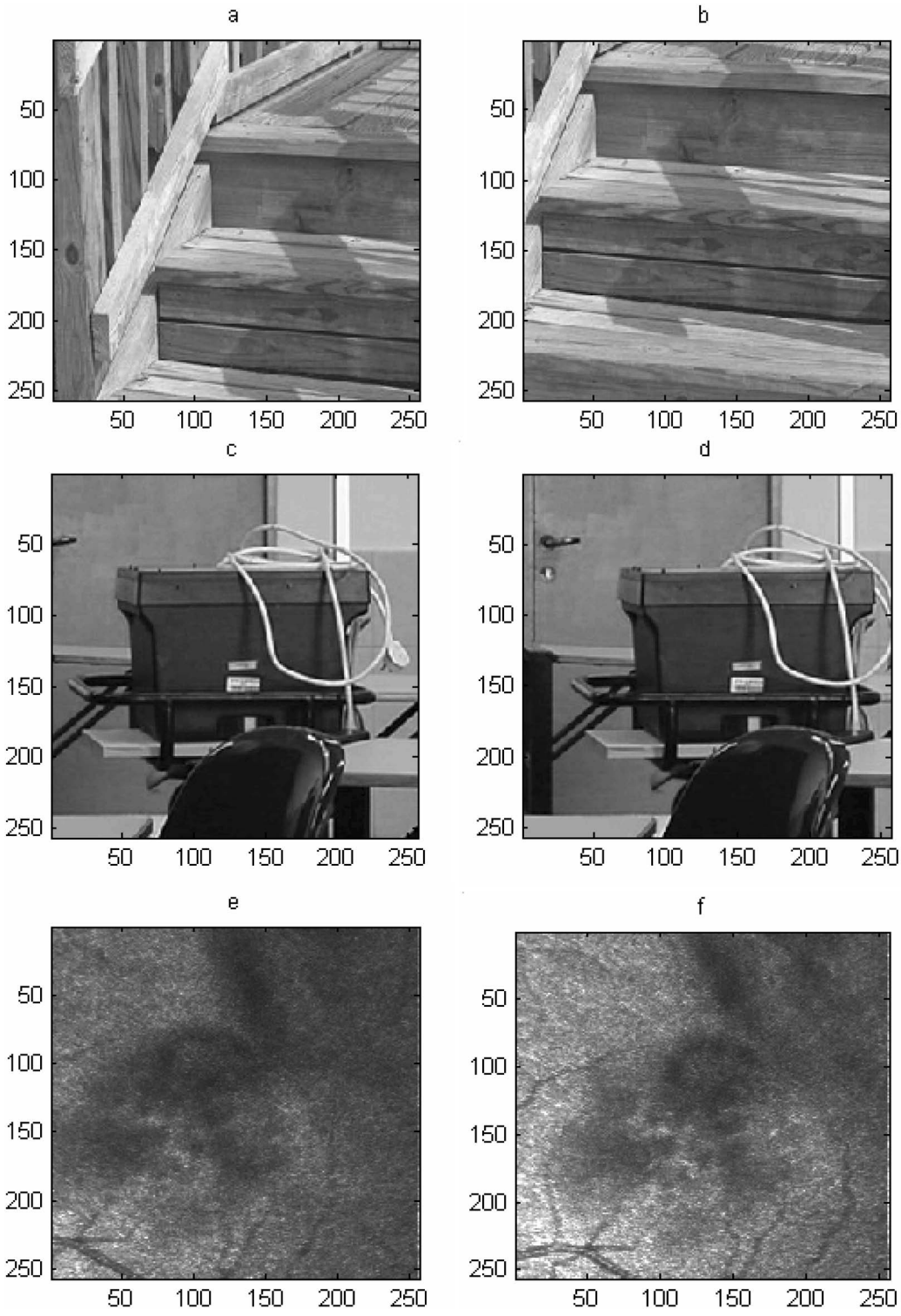


Fig. 2 Original input images used in the demonstration (256×256 pixels, with 4-deg visual-angle span assumed): (a) and (b) images shifted digitally; (c) and (d) images shifted as a result of a camera-viewpoint change; (e) and (f) very noisy nonidentical vessel images recorded at different times.

4.1 *Vision-Model-Based Feature Detection Method (Space Invariant)*

The input image is first bandpass filtered through a multi-scale set of visual channels according to Eq. (1). The bandpass-filtered images (scales) are then thresholded according to the foveal CT of the visual system. For each channel i, the threshold value $Th_{0,i}$ at its spatial center frequency 2^i is determined (from human psychophysical foveal data),[40] and applied to the filtered image in the spatial domain $S_i(x,y)$ as follows

$$T_i(x,y)=\begin{cases}+1, & \text{if } S_i(x,y)\geqslant +Th_{0,i}\\ 0, & \text{if } -Th_i<S_i(x,y)<+Th_{0,i},\\ -1, & \text{if } S_i(x,y)\leqslant -Th_{0,i}\end{cases} \quad (5)$$

producing a trilevel thresholded i'th image $T_i(x,y)$ for each scale. Detection results are coarser as the channel spatial frequency becomes lower. A trilevel image E of detected visual features is then obtained, based on the correspondence across all scales. This stage applies an all-or-none decision, carried out using the visual information derived in the previous stage.

$$E(x,y)=\begin{cases}+1, & \text{if } T_i(x,y)=+1,\ \forall\, i\in\Re\\ 0, & \text{otherwise}\\ -1, & \text{if } T_i(x,y)=-1,\ \forall\, i\in\Re\end{cases}, \quad (6)$$

where $\Re$ is the set of channels used (usually the four highest channels). The resulting detection differs from conventional feature (edge) detection techniques in that the edge features are represented in pixel pairs, with a black pixel located on the darker side of the edge and a white pixel located on its brighter side. Bar features are represented by single pixels of the correct polarity.

4.2 *Extension to Different Eccentricities*

When processing visual information within a wide visual field, the space variance characteristics should be taken into account. The feature detection technique employs several filters that correspond to different visual channels. For each eccentricity, only channels that transfer significant signal quantities should be employed in the feature detection process. Channels with zero output in Table 2 obviously cannot be used for the feature detection task. The minimal energy value that allows the use of a channel can be found experimentally. Since higher orders provide the finer details, it is preferable to use the highest order channels as possible, limited by the requirement for significant intensity transfer by the channel. Lower orders are required mostly to reduce the noisiness of the highest order information and contribute minimally to feature detection. The resulting feature detection at eccentricity θ [modified Eq. (6)] becomes:

$$E_\theta(x,y)=\begin{cases}+1, & \text{if } T_{\theta,i}(x,y)=+1,\ \forall\, i\in\Re_\theta\\ 0, & \text{otherwise}\\ -1, & \text{if } T_{\theta,i}(x,y)=-1,\ \forall\, i\in\Re_\theta\end{cases}, \quad (7)$$

where the set of channels used, $\Re_\theta$, is determined according to the signal quantities transferred by them, as shown for example in Table 2. The number of channels in the fovea that we included in $\Re$ was four.[23,28] This number may be reduced with increases in eccentricity, because fewer active channels are available there (Table 2) and the low frequency channels have a limited role in feature detection.

5 Motion Estimation from Foveated Images

Motion (displacement) between foveated images with different AOIs (foveal locations) is estimated here by both block-matching and phase-correlation image registration techniques. Registration is applied and compared using two types of visual information: foveated images directly, and foveated feature-detected images.

5.1 *Digital Image Registration*

To put the visual-information-based registration performed here in a conventional context of image registration, we briefly review the registration process, using four basic components[41]: the *feature space* is the information from the images used to perform the matching between the images (for example, the pixels grayscale values, the feature locations); the *search space* is the set of potential transformations that establish the correspondence between the images (for example, shift, rotation, rescaling); the *search strategy* decides how one chooses the next step (for example, orderly raster, random selection); and the *similarity metric* determines the match measure between one image and the other image (for example, absolute differences, correlation, mean squares). In this work, visual information (either a grayscale image or features detected in this image) is used as a feature space, and the search space is all the possible shifts between the images within an assumed maximum displacement. For the purpose of motion estimation from foveated images, we examine phase-correlation[42] and block-matching registration techniques. In block matching, the sum of absolute values of differences (SAVD)[43] and the mean-square error (MSE) are used as metrics. The difficulty in registration results from the fact that the displaced images have space-variant (retinal-like) resolution with different foveal locations, which means that the matched objects are not identical (different levels of blurriness).

Phase correlation makes use of the shift property of the Fourier transform, in which a shift in the spatial domain causes a phase shift in the frequency domain. Given two displaced (originally identical) images,

$$f_a(m,n)=f_b(m-m_o,n-n_o), \quad (8)$$

their corresponding Fourier transforms will be related by

$$F_a(w_m,w_n)=F_b(w_m,w_n)\exp -j(w_m m_o+w_n n_o). \quad (9)$$

The inverse Fourier transform of the normalized cross-power spectrum of the two images will be:

$$PC(m,n)=\Im^{-1}\left[\frac{F_a(w_m,w_n)\cdot F_b(w_m,w_n)^*}{\left|F_a(w_m,w_n)\cdot F_b(w_m,w_n)\right|}\right], \quad (10)$$

where F_a and F_b are the Fourier transforms of the displaced images, $*$ denotes the complex conjugate, and $\Im^{-1}$ symbolizes the inverse Fourier transform. The numerator in the brackets is the Fourier transform of the cross correlation. The denominator in the brackets performs as a whitening (normally high-pass) filter. Ideally, the resulting $PC(m,n)$ is a delta function located at the displacement (m_o,n_o). However, in the case of displaced foveated images, Eq. (8) is not accurate, because displaced features have different resolution (different blurriness) as a result of the change of the foveal location (the AOI).

In the block-matching registration process, a subsection (*template*) in one image is compared to shifted subsections with the same size throughout a search area in the second image, according to the similarity measure. The shift that

gives the best match is the registration result. Enabling search movements in all directions, the normalized SAVD between a template A of size $J \times K$ and a shifted subsection B of the same size in the second image is[43]:

$$\mathrm{SAVD}(m,n) = \frac{1}{J \cdot K} \sum_{j=-J/2}^{J/2-1} \sum_{k=-K/2}^{K/2-1} \left| (A_{jk} - \bar{A}_{jk}) - (B_{j-m,k-n} - \bar{B}_{j-m,k-n}) \right|, \tag{11}$$

where m and n are the horizontal and vertical shifts in the search areas, and $\bar{A}_{jk}$ and $\bar{B}_{j-m,k-n}$ are the averages of the gray levels around A_{jk} and $B_{j-m,k-n}$, respectively. The subtraction of the averages normalizes the subsections to have zero average, and improves the reliability of the registration, since it decreases the effect of different average luminance levels between the images.[43] The registration point is at the values of m and n, where the SAVD has a minimum value.

The normalized MSE metric is defined as:

$$\mathrm{MSE}(m,n) = \frac{1}{J \cdot K} \sum_{j=-J/2}^{J/2-1} \sum_{k=-K/2}^{K/2-1} \left[(A_{jk} - \bar{A}_{jk}) - (B_{j-m,k-n} - \bar{B}_{j-m,k-n}) \right]^2. \tag{12}$$

To reduce the computation load, such metrics can be implemented with a threshold.[43] In this case, the accumulation process in Eqs. (11) and (12) is stopped when the error obtained by the subtraction exceeds a threshold level, reducing the tests for many possible registration points (m,n) with high error levels. Features matching may be particularly useful in the context of such thresholded applications of the similarity measure.

6 Results

Examples of the results are presented here for image foveation (described in Sec. 3), feature detection in foveated images (described in Sec. 4), and estimation of motion between foveated images (described in Sec. 5). In the first example (digital shift), images with added noise were arbitrarily shifted apart to demonstrate results with accurately known ground truth. To examine a more realistic case where images are not shifted by an integer number of pixels, in the second example (real shift) images are shifted apart as a result of two different viewpoints of a digital camera. The input images used in the examples are presented in Fig. 2. The third example uses noisy retinal-vessel image frames taken from the same video sequence, obtained with a scanning laser ophthalmoscope. In this case, the displaced input images (due to the natural movement of the recorded eye) are also spatially distorted, one with respect to the other, due to the system's optics, as can be observed in Figs. 2(e) and 2(f) For comparison purposes, all metrics were applied to all the cases. The result of the phase correlation technique is shown in the first example. Block-matching results are shown in the second and third examples, with MSE and SAVD metrics, respectively.

6.1 *Digital Simulated Image-Shift Example*

Two input images shifted horizontally and vertically $(-50,-60)$ pixels apart, as shown in Figs. 2(a) and 2(b), with added noise forming 10-dB SNR, were foveated as described in Sec. 3, at an arbitrary AOI (120, 120) in both images. The shifted foveated images are presented in Figs. 3(a) and 3(c). Space-variable (foveated) feature detection was performed for each shifted image according to Sec. 4. Channel sets $\Re_\theta$ for different eccentricities used in the feature detection procedure [Eq. (7)] were selected, as shown in Table 2. (The channels used for each eccentricity are marked by bold numerals.) The feature detection results are shown in Figs. 3(b) and 3(d). It can be seen that detected features become coarser as eccentricity increases and the foveated image resolution decreases. The result of frame-to-frame motion estimation, extracted by registration of the shifted foveated images, is shown in Fig. 3(e). Motion estimation from the feature-detected images is shown in Fig. 3(f). In both cases, the phase-correlation technique was applied. Both the grayscale and the feature images resulted in the identical and correct registration at the same shift of $(-50,-60)$. The same motion displacement was also found with the two other metrics. Experiments have been done with 22 different images, and high noise levels added to both the original input and the foveated images, producing (−)10-dB SNR and higher. Accurate registration was obtained for noise levels producing 0-dB SNR and higher, with both phase-correlation and block-matching techniques.

6.2 *Real Shift Example*

This example [with the shifted images shown in Figs. 2(c) and 2(d) as input] is presented in Fig. 4. In this more realistic case, the shift between the images was created by changing the viewpoint of the digital camera that acquired the images. This results in a shift that is not necessarily an integer number of pixels, and in two images that have slightly different content in addition to the shift, due to noise and different integration across the pixels of the sensor. Image foveation, feature detection, and motion estimation were implemented in a manner similar to the previous example, and are presented in Fig. 4 (the channels used were also similar). The registration maps in Figs. 4(e) and 4(f) are the normalized MSEs [Eq. (12)] at every shift distance within a 100×100 pixel search area that is the assumed maximum possible displacement between the images (±50 pixels in the horizontal and vertical directions). In this case, a darker pixel represents better similarity between the matched regions. Both registration maps show a minimum at a (22, 0) pixel shift. The identified shift was verified by visually comparing magnified object locations (edges, corners) in the manually shifted input images. The registration map of the edge-detected features has a significantly sharper peak at the registration point, but it is less smooth than the registration map based on the intensity images. This sharpness property of the registration point relative to its surrounding is expected when thin features such as object contours are used as a feature space. The bipolar nature of the visual-model-based detected features used here can further increase the relative sharpness of the

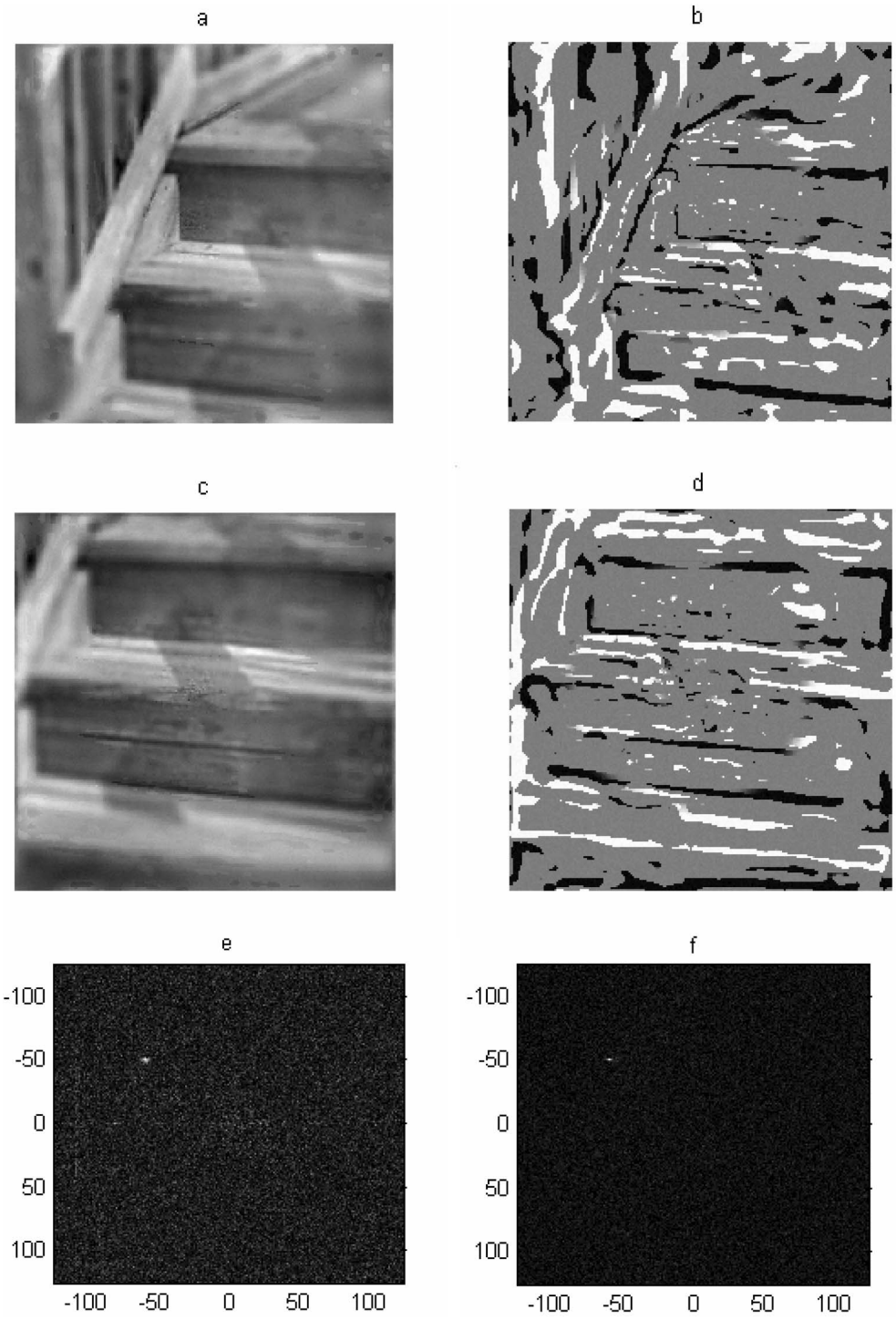


Fig. 3 Results for the digitally shifted images of Figs. 2(a) and 2(b): (a) and (c) images foveated at (120, 120) and shifted (−50, −60) pixels apart; (b) and (d) feature detection results for the foveated images of (a) and (c), respectively; (e) and (f) the phase correlation results produced from the foveated images and features, respectively. The correct shift is identified by the highest (brightest) point in each map.

registration point. This means that when features are used, the errors away from the true registration point (represented by the brightness of the pixels around the point) are bigger on average. This can reduce the computation load when a threshold is used (as described at the end of Sec. 5), because these points can reach the threshold and be dropped faster. The same displacement results were produced by the two other metrics. Here too, the registration carried out with the feature maps provides a sharper peak than that carried out with the grayscale images.

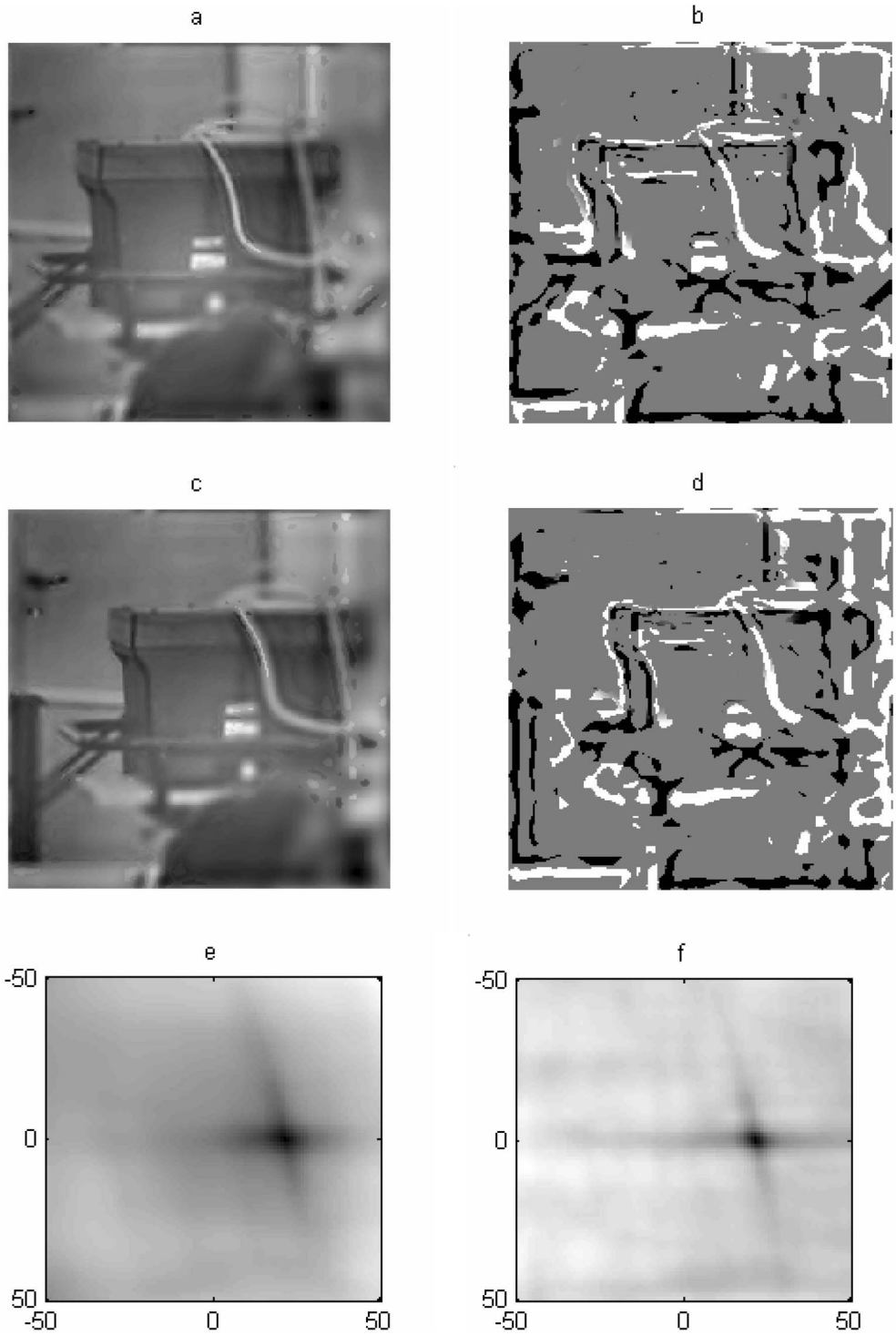


Fig. 4 (a), (b), (c), and (d) are the same as in Fig. 3, but for images shifted by a change of the camera viewpoint; (e) and (f) are the MSE maps produced from the foveated images and features, respectively. The correct shift is identified by the lowest (darkest) point in each map. The shift of (22, 0) as identified from both maps was verified by manually comparing magnified object locations in the input images.

6.3 *Extremely Noisy and Nonidentical Image Example*

In this case, presented in Fig. 5, retinal images taken by a scanning laser ophthalmoscope [shown in Figs. 2(e) and 2(f)] were used. The images are very noisy and the objects in the images are not identical and have different brightness.

The foveated images are shown in Figs. 5(a) and 5(c), and the foveated features are shown in Figs. 5(b) and 5(d). The SAVD metric identified a displacement of (27, 6) pix-

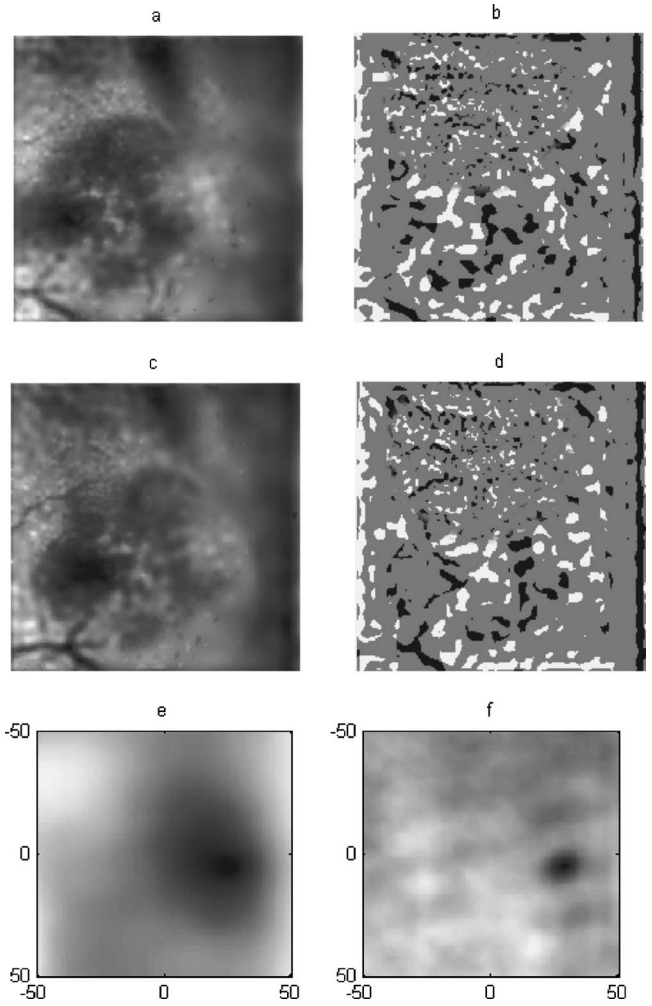


Fig. 5 Same as Fig. 4, but for the noisy nonidentical vessel images. The identified shifts according to the SAVD registration maps (e and f) were (27, 6) pixels and (29, 6), respectively. A (29, 6) pixels shift was estimated manually.

els between the images when the foveated images were used in the registration process [Fig. 5(e)], and (29, 6) pixels when the features were used [Fig. 5(f)]. The MSE metric identified a displacement of (26, 6) pixels when the foveated images were used, and (29, 6) pixels when the features were used. Manual comparison using magnified salient locations was not easy to accomplish because of the noise, nevertheless it was judged that the correct displacement was about (29, 6) pixels. Since the manual registration also used the features only, it tends to agree with the feature registration, but this is probably more accurate than using the overall gray level on the fairly uniform but noisy retinal background.

7 Summary and Conclusions

This work presents a vision-model-based method to image foveation and feature detection in foveated images. Foveated images are created using the visual system's CT, which varies with eccentricity and the spatial frequency of the visual channels. A systems-engineering approach is employed as a general tool to quantitatively analyze the multichannel vision model. A new notion—the channel modulation transfer function area (CMTFA)—that evaluates the ability of a single visual channel to transfer image contrast is defined, and image content transferred at each channel is produced. This information is applied to select which set of channels could be used in a feature-detection procedure extended for implementation in a foveated system. With the systems-engineering approach, incorporation of another system or component (between the scene and the observer) simply involves multiplying the additional system's MTF by the existing system response [$C_i(r)$ of the visual system]. This would change the quantities shown in Tables 1 and 2. If, for example, the quantities in Table 2 are reduced at the higher channels (which is the expected effect of any real imaging system, where the MTF decreases at higher frequencies), the foveated image will be blurrier and the channels selected for the feature-detection process may change.

Estimation of the motion displacement between foveated images is performed by image registration, using the shifted foveated images and the shifted detected-feature maps. The methodology developed here corresponds directly to visual processing according to contemporary vision models. The high efficiency of the visual process, and its compatibility to the human eye that may be the end user of the foveated images, argues for using such vision-based imaging systems. Registration techniques examined for motion estimation include phase correlation and block matching (with MSE and SAVD metrics), and usually gave similar results. Accurate results are obtained for high levels of added noise (down to 0-dB SNR), in spite of the change of feature resolution that results from the change of the AOI (foveal location). When comparing motion estimation results obtained with detected edge features versus foveated images, we can see that when features are used, the identified displacement point is more distinct from its surroundings, which may result in lower computational loads. The use of detected features produces slightly better results in cases where the displaced images are taken at different conditions (e.g., point of view and lighting), and in high noise conditions. Noisy homogeneous areas in the image, which may have damaging contributions in the registration process, are removed during feature detection, which preserves locations of brightness transitions.

Acknowledgments

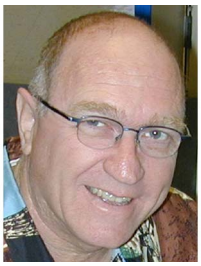
Supported in part by NIH grants EY05957 and EY12890.

References

1. W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," *Proc. SPIE* **3299**, 294–305 (1998).
2. S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia* **4**(1), 129–132 (2002).
3. E. L. Schwartz, "Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding," *Vision Res.* **20**(8), 645–669 (1980).
4. R. Wallace, P. W. Ong, B. Bederson, and E. Schwartz, "Space variant image processing," *Int. J. Comput. Vis.* **13**(1), 71–90 (1994).
5. E. C. Chang and C. K. Yap, "A wavelet approach to foveating images," in *13th ACM Symp. Computational Geometry*, pp. 397–399 (1997).
6. W. S. Geisler and J. S. Perry, "Variable-resolution displays for visual communication and simulation," *Soc. Info. Display* **30**, 420–423 (1999).
7. S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compres-

sion with optimal rate control," *IEEE Trans. Image Process.* **10**, 977–992 (2001).
8. D. F. Wilson and J. P. Siebert, "Foveated stereo for an active robot vision system," *European Conf. Visual Perception (ECVP), Perception* **22** Supplement, 114–115, Pion Ltd, London (1993).
9. A. Basu and K. J. Wiebe, "Enhancing videoconferencing using spatially varying sensing," *IEEE Trans. Syst. Man Cybern.—Part A: Systems and Humans* **28**(2), 137–148 (1998).
10. M. L. Thomas, R. Robinson, W. P. Siegmund, and S. E. Antos, "Fiberoptic development for use on the fiber optic helmet-mounted display," *Opt. Eng.* **29**, 855–862 (1990).
11. P. T. Kortum and W. S. Geisler, "Implementation of a foveated image-coding system for bandwidth reduction of video images," *Proc. SPIE* **2657**, 350–360 (1996).
12. H. D. Warner, G. L. Serfoss, and D. C. Hubbard, "Effects of area-of-interest display characteristics on visual search performance and head movements in simulated low-level flight," Armstrong Laboratory, Human Resources Directorate, Aircrew Training Division, Williams AFB, AZ (1993).
13. C. Stiller and J. Konrad, "Estimating motions in image sequences—a tutorial on modeling and computation of 2D motion," *IEEE Signal Process. Mag.* **16**, 70–91 (1999).
14. B. Bridgeman, A. H. C. van der Heijden, and B. Velichkovsky, "Visual stability and saccadic eye movements," *Behav. Brain Sci.* **17**(2), 247–258 (1994).
15. D. Heiner, B. Bridgman, and W. X. Schneider, "Immediate post-saccadic information mediates space constancy," *Vision Res.* **38**, 3147–3159 (1998).
16. D. O. Bahcall and E. Kowler, "Illusory shifts in visual direction accompany adaptation of saccadic eye movement," *Nature (London)* **400**, 864–866 (1999).
17. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman and Company, New York (1982).
18. H. Spitzer, M. Almon, and I. Sherman, "A model for the early stages of motion processing based on spatial and temporal edge detection by X-cells," *Spatial Vis.* **8**, 341–368 (1994).
19. D. C. Burr, M. C. Morrone, and D. Spinelli, "Evidence for edge and bar detectors in human vision," *Vision Res.* **29**, 419–431 (1989).
20. E. Peli, R. A. Augliere, and G. T. Timberlake, "Feature-based registration of retinal images," *IEEE Trans. Med. Imaging* **6**(3), 272–278 (1987).
21. H. Li, B. S. Manjunath, and S. K. Mitra, "A contour-based approach to multisensor image registration," *IEEE Trans. Image Process.* **4**(3), 320–334 (1995).
22. N. S. Kopeika, *A System Engineering Approach to Imaging*, SPIE Optical Engineering Press, Bellingham, WA (1998).
23. H. R. Wilson and J. R. Bergen, "A four mechanism model for threshold in spatial vision," *Vision Res.* **19**, 19–32 (1979).
24. A. B. Watson, "Summation of grating patches indicates many types of detector at one retinal location," *Vision Res.* **22**, 17–25 (1982).
25. N. Graham, J G. Robson, and J. Nachmias, "Grating summation in fovea and periphery," *Vision Res.* **18**(7), 815–825 (1978).
26. M. A. Garcia-Perez and V. Sierra-Vazquez, "Do channels shift their tuning towards lower spatial frequencies in the periphery?" *Vision Res.* **36**(20), 3339–3372 (1996).
27. C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications," in the *Proc. Intl. Conf. Acoustics, Speech, Signal Process.*, pp. 2293–2296, (1996).
28. E. Peli, "Feature detection algorithm based on a visual system model," *Proc. IEEE* **90**, 78–93 (2002).
29. J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters," *J. Opt. Soc. Am. A* **2**(7), 1160–1168 (1985).
30. E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A* **7**(10), 2032–2040 (1990).
31. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
32. E. Peli, J. Yang, and R. B. Goldstein, "Image invariance with changes in size: The role of peripheral contrast thresholds," *J. Opt. Soc. Am. A* **8**(11), 1762–1774 (1991).
33. M. A. Georgeson and G. D. Sullivan, "Contrast constancy: Deblurring in human vision by spatial frequency channels," *J. Physiol. (London)* **252**(3), 627–656 (1975).
34. J. G. Robson and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vision Res.* **21**, 409–418 (1981).
35. M. W. Cannon Jr., "Perceived contrast in the fovea and periphery," *J. Opt. Soc. Am. A* **2**, 1760–1768 (1985).
36. J. S. Pointer and R. F. Hess, "The contrast sensitivity gradient across the human visual field: With emphasis on the low spatial frequency range," *Vision Res.* **29**, 1133–1151 (1989).
37. E. Peli and G. A. Geri, "Discrimination of wide-field images as a test of a peripheral-vision model," *J. Opt. Soc. Am. A* **18**, 294–301 (2001).
38. G. J. P. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, SPIE Optical Engineering Press, Bellingham, WA (1999).
39. G. J. Burton and R. Moorhead, "Color and spatial structure in natural scenes," *Appl. Opt.* **26**(1), 157–170 (1987).
40. E. Peli, L. Arend, G. Young, and R. Goldstein, "Contrast sensitivity to patch stimuli: Effects of spatial bandwidth and temporal presentation," *Spatial Vis.* **7**(1), 1–14 (1993).
41. L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.* **24**(4), 325–376 (1992).
42. C. D. Kuglin and D. C. Hines, "The phase correlation image alignment method," *IEEE Conf. Cybernetics Soc.*, pp. 163–165 (1975).
43. D. I. Barnea and H. F. Silverman, "A class of algorithms for fast digital image registration," *IEEE Trans. Comput.* **C-21**, 179–186 (1972).

Yitzhak Yitzhaky received his BS, MS, and PhD degrees in electrical and computer engineering from Ben Gurion University, Israel, in 1993, 1995, and 2000, respectively. From 2000 to 2002, he was a postdoctoral research fellow at the Schepens Eye Research Institute, Harvard Medical School, Boston, Massachusetts. Currently, he is with the Electro-Optics Unit at Ben Gurion University. His research is mainly in the fields of image restoration, image enhancement, and computer vision.



Eli Peli received a BSEE, cum laude, and MSEE from the Technion - Israel Institute of Technology. He earned his doctorate from the New England College of Optometry in Boston. He is Senior Scientist and the Moakley Scholar in aging eye research at the Schepens Eye Research Institute, and professor of ophthalmology at Harvard Medical School. He also serves on the faculty of the New England College of Optometry (adjunct professor of optometry and visual sciences) and the Tufts University School of Medicine (adjunct professor of ophthalmology). He is the director of the Vision Rehabilitation Service at the New England Medical Center Hospitals in Boston. He is a Fellow of the American Academy of Optometry, the Optical Society of America, and the Society for Information Display (SID). His principal research interests are image processing in relation to visual function, and clinical psychophysics in low vision rehabilitation, image understanding, and evaluation of display-vision interaction. He has published more than 100 scientific papers and has been awarded six U.S. Patents.