Face Reconstruction Using Fixation Positions and Foveated Imaging

Fang Fang[1,2] Zhiguo Ma[2] Laiyun Qing[3,2] Jun Miao[2] Xilin Chen[2] Wen Gao[4,1]
[1] School of Computer Science and Technology, Harbin Institute of Technology, China
[2] Key Lab of Intelligent Information Processing of CAS,
Institute of Computing Technology, CAS, Beijing 100190, China
[3] Graduate School of the Chinese Academy of Sciences, CAS, Beijing 100039, China
[4]Institute for Digital Media, Peking University, China
{ffang, zgma, lyqing, jmiao, xlchen, wgao}@jdl.ac.cn

Abstract

Face representation is important for face recognition system. Though most popular face representations are based on uniform grid sampling, some recent face recognition systems adopt weighted sampling on the different regions of a face. Psychological analysis of visual attention or human eye fixations on human face images may suggest some cues for face representation. Human Visual System(HVS) gives different weight to different region of human face via space-variant sampling on fovea and non-uniform distribution of fixations. This paper focuses on the problem of simulation of the foveated imaging phenomenon in HVS, and introduction of foveated imaging method into reconstruction of face in region of interest (ROI) using different fixation sources. We compare the effectiveness of actual fixation on reconstruction of face in ROI with uniform, random distribution fixation as well as fixation generated by artificial model. The experimental results on 100 face images from FRGC [7] data set show that actual fixation positions and model-generated fixation positions reconstruct the face in ROI with considerably better quality. A further analysis on the statistics of fixation positions also shows that the distribution of the fixation points is consistent with the weights of different regions on face images used in some other face recognition systems.

1. INTRODUCTION

More than 80% information attained to human beings is through visual perception. Visual attention is an important component of Human Visual System (HVS). It plays a fundamental role in understanding scenes by sequentially searching the most informative parts of image [17] with discrete fixations linked by saccadic eye movements. This strategy eases the need of real-time perceiving of the global information contained in scenes. Moreover, HVS doesn't weigh every pixel in the field of view with the same importance — it is highly space-variant in sampling, coding, processing and understanding as the spatial resolution of the fovea is highest around the point of fixation (foveation point) and decreases rapidly with increasing eccentricity [4]. By taking the advantage of this fact, it is possible to selectively reconstruct objects in ROI with considerable better quality.

Many researchers have adopted foveated imaging method in vision-related tasks, such as real-time perceptually lossless low-bandwidth video communication [4], video compression [9], and visual search for corners [2].

In recent years, with emerging interest in active vision [1], computer vision researchers attempted to build models of attention mechanism, which was later applied in the field of image perception and identification. Consequently, a number of computational attention models were developed, such as the models proposed in [13, 15]. The basic principles behind these efforts were greatly influenced by psychophysical research. Based on the work in [3, 11, 12], Itti proposed a saliency-based visual attention model for scene analysis in [10]. In this work, visual input was first decomposed into a set of topographic feature maps which all feed into a master 'saliency map' in a bottom-up manner.

Rybak et al. [14] proposed a model which contains motor control directives stored in a 'where' memory and locally expected bottom-up features stored in a 'what' memory. This model used eye movement scan-paths as sensor-motor memories for recognition. It can recognize complex gray-scale scenes and faces in a translation-, rotation- and scale-independent manner. In this paper we have implemented artificial model to generate gaze positions and compare the performance between uniform, random, eye tracker recorded and artificial model generated gaze positions.

Our contributions in this paper include: (1)simulate the foveated imaging phenomenon with a simple yet flexible

978-1-4244-2154-1/08/$25.00 ©2008 IE

foveated imaging method, (2) reconstruction of the face in ROI using different fixation source, and comparing the effectiveness of uniform, random distribution fixations with model-generated, actual fixations recorded by eye tracker device, and (3) statistical analysis of the distribution of fixation positions acquired on 4 subjects to verify that human put different weights on different regions of face images.

This paper mainly focuses on simulation of foveated imaging phenomenon of HVS using a simple yet flexible foveated imaging method, and compare the reconstruction performance on different fixation sources using our foveated imaging method. The rest of the paper is organized as follows. The detail of our foveated imaging method and artificial fixation model inspired by Rybak [14] is described in Section 2. Experimental results are shown in Section 3. The conclusion and discussion are given in Section 4.

2. FOVEATED IMAGING METHOD AND ARTIFICIAL FIXATION MODEL

Researchers simulate the foveated imaging phenomenon with different methods. Earlier systems used special purpose hardware and create foveated image by increasing pixel-element size as a function of angular distance from the point of fixations. More recently, Silsbee, Bovik and Chen [16] implemented a computation effective foveated block pattern matching algorithm. Geisler and Perry [4] proposed a multi resolution pyramid coding method and the corresponding motion estimation algorithms for real-time low bandwidth video communications. In this paper we employ a simple yet flexible approach similar with [4] to simulate the foveated imaging phenomenon.

2.1. Foveated imaging method

We simulate the foveated imaging phenomenon in HVS with different Gaussian smooth operator applied in different regions of image. The Gaussian smooth operators with increasing window widths and variances are applied to several circles which double its radii from the point of the fixation.

The 2D form of Gaussian function is:

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

where (x,y) is the coordinate of the current pixel, σ^2 is the variance of the Gaussian function. Taking the discrete form of 2D Gaussian function within small window, and convolution it with the original image, a blurred image will be generated.

A sample image generated by our foveated imaging method with only one fixation at the body of the bee is shown in Figure 1.

When reconstructing an image, (See Figure 2) we perform following steps on every pixel of the image: (1) compute the distances between current pixel and all the fixations, (2) find the closest fixation from the current pixel, (3) select corresponding window width and variance of Gaussian smooth operator according to the distance between current pixel and its closest fixation, (4) employ the gaussian smooth operator with known window width and variance on original image, (5) and select the pixel value of the blurred image with the same coordinate as the value of the current pixel.

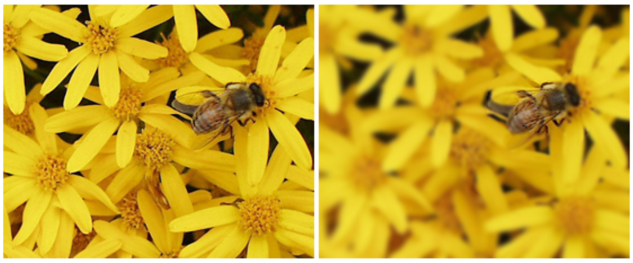

Figure 1. Original image(left) and its foveated image(right). Original image from the supplementary material of [5] , and the foveated image is generated by our foveated imaging method.

To avoid the repeatedly use of the Gaussian smooth operator on the original image, we employ the Gaussian smooth operator on the original image with different window width and variance and store the blurred image for future use. When we want to find the value of the current pixel, we select the corresponding blurred image according to the closest distance between the current pixel and its closest fixations, and select the value of the pixel at same place on the blurred image as the value of the current pixel.

2.2. Artificial model for fixation positions generation

Rybak model [14] is developed on the base of biologically plausible algorithms and has demonstrated the ability to recognize complex grey-level images, like human face images. We adopt Rybak model to generate fixation points with following steps (See Figure 3):

1) Image transformation: convert color image to grey-level image and scale the grey-level image with maximum of its width and height equal to 128 pixels (Figure 3a).

2) Primary transformation: To simulate the decrease in the resolution of visual field perception from the fovea to the retinal periphery; The attention window (AW) performs a primary transformation of the image into a ‘retinal image’ at the fixation point. The AW consists of three rings with increasing radius and decreasing resolution(Figure 3b).

3) Extraction of primary features: The retinal image in the AW is used as input to the module for extraction of primary features which performs a function similar to the primary visual cortex. Orientation tuning of a neuron was determined by its receptive field which was formed as the

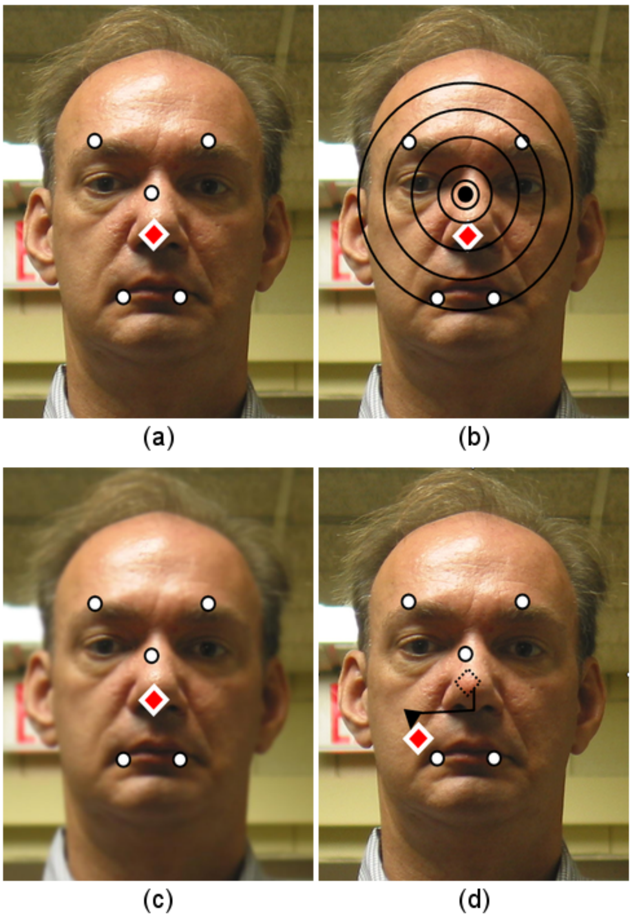

Figure 2. Illustration of reconstruction steps: (a) Sample image with 5 fixations (denoted as circles) and the current pixel (denoted as the red diamond) , (b) The closest fixation of current pixel (the solid one) and the circles center at the current pixel, (c) Result image after employing the Gaussian smooth operator, and (d) move to next pixel, then apply the above steps on new pixel again.

difference of two Gauss convolutions with spatially shifted centers. Sixteen neurons tuned to different edge orientations (different directions of brightness gradients) interacted competitively owing to the strong reciprocal inhibiting connections, the orientation tuning of the winning neuron determined the edge orientation in the given point. The step of orientation -22.5° was taken as the unit of the angle measure. In each fixation, the oriented edges were extracted in the fixation point(the basic edge segment) and in 48 context points lying on intersections of 16 radiating lines differing in 22.5° and of concentric circles with exponentially increasing radii(Figure 3c). The oriented edges corresponding to the first circle were extracted with the same resolution as the basic one. The resolutions with which the other edges were extracted were determined by their position in the attention window.

4) Selection and representation of the next fixation point: Use the following formula to select next fixation point:

$$A_k = a_1 \cdot \frac{Z_k}{Z_{max}} + a_2 \cdot \frac{\lambda_k}{2} + a_3 \cdot \eta_{ij}(n) + a_4 \cdot \chi_{ij} \quad (2)$$

where the first term determines the normalized value of brightness gradient in the context point, which is defined by the output of the corresponding neuron-winner in the module for primary feature detection. The second term determines the relative distance of the context edge from the center of the retinal image. The third term is incorporated to prevent 'cycling', the function $\eta_{ij}(n)$ determines a 'novelty' of the vicinity of the context point. The fourth term predefines a 'semantic significance' of the area. In this paper, the region of face is described as 'semantic significance'. The scan path and fixation points of image viewing on the background of the initial image are show in Figure 3d.

3. EXPERIMENTAL RESULTS

We use EyeLink II head mounted eye tracking system [6] to record fixation positions, which has a data sampling rate of 500Hz when recording binocular fixation data and an average fixation error less than 0.5 degree.

In our experiments, four male colleague students who are naive to the purpose of the experiments are selected as subjects. Each image is presented to subject at the center of a 19 inch high refresh rate CRT monitor which has the resolution of 1600*1200. Subjects sit in the front of the monitor with a distance about 60cm (which corresponds to approximately 36*27 degrees of visual angles, equivalent to 44 pixels per visual degree) with a chin rest to help keep the head of the subjects as still as possible.

After the calibration process, we verify the tracking performance by performing a test tracking with 9 known points and select the dominant eye with smaller average tracking error. We perform a drift correction before each image is shown to subjects by displaying a small target centered at the monitor screen and adjust the drift error to avoid fixation drift caused by head movement of subjects. We present each image to subjects for 5 seconds and record the eye fixations of subjects.

In our experiments we use 5 circles with the radius of the most inner circle equal to 12 pixels and the smoothing window width of the most inner circle equal to 1 (guarantee no blur inside the most inner circle) to simulate the foveated imaging system. The radii of outer circle are 24, 48, 96, 192, respectively, and the corresponding smooth window width are 5, 9, 13, 17, respectively. The variance of a Gaussian smooth operator is selected as a function of Gaussian window width:

$$Variance(n) = (n/2 - 1) * 0.3 + 0.5 \quad (3)$$

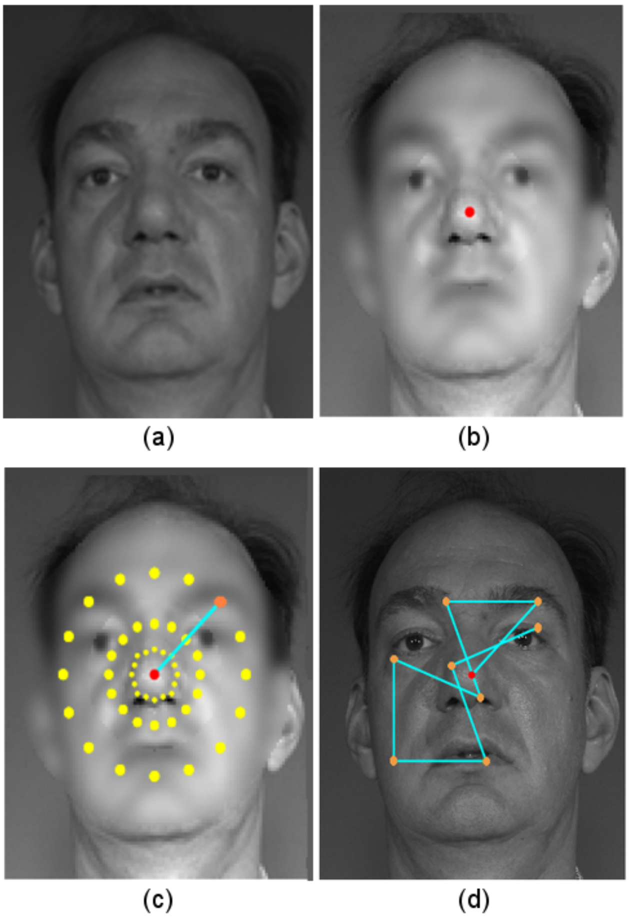

Figure 3. (a) The initial image, (b) Image transformation within the AW in one fixation point, (c) The 48 context points (show by the yellow point), the central red point denotes the current fixation point, the yellow point denotes the next fixation point, (d) The scan path and fixation points of image viewing on the background of the initial image.

where n is the window width (pixels). This is the default setting of OpenCV [8] Gaussian smooth operator for faster computation on small convolution kernels.

Some fixations recorded by the eye tracker device tend to be close with each other. We purge the number of fixation to 9 by repeatedly remove one of two closest fixations. Benefit from the careful crop of the images, we can simply define the face region or ROI as the central part of the image with 60% width and 50% height.

We use the histogram entropy and the root mean square error(RMSE) in face region to verify the validity of our reconstruction method. The histogram entropy is a function of the probability of each grey-level in histogram bins:

$$Entropy(p) = \sum_{i=1}^{256} -p_i * \log_2(p_i) \tag{4}$$

Table 1. The RMSE and entropy when using different fixation sources , fixations data from 4 subjects, 2 random, 1 uniform, and artificial model.

	fixations	RMSE		Entropy	
	sources	avg	std dev	avg	std dev
Subject	subject 1	3.84	1.32	7.257534	0.203520
	subject 2	4.00	1.34	7.257657	0.203442
	subject 3	4.05	1.40	7.257851	0.203426
	subject 4	3.90	1.35	7.258114	0.203556
Random	Random 1	4.48	1.52	7.257694	0.203923
	Random 2	4.46	1.50	7.257649	0.203798
Uniform	uniform	4.44	1.52	7.258869	0.204285
Model	Model	3.75	1.32	7.295988	0.199254

where p_i is the probability of i-th grey-level in the histogram bins from 1 to 256.

The root mean square error(RMSE) between the original image and the reconstructed image in ROI is also be used to verify the effectiveness of our method. The RMSE in ROI is a function of the pixel values of the original and the reconstructed image in ROI:

$$RMSE = \sqrt{\frac{1}{w*h*c}\sum_{i=1}^{w}\sum_{j=1}^{h}\sum_{k=1}^{c}(O(i,j,c)-R(i,j,c))^2} \tag{5}$$

where $(i,j) \in$ ROI , w is the width of the image, h is the height of the image, c is the number of channels of the image, $O(i,j,c)$ is the pixel value of the original image on channel c at position (i,j), and $R(i,j,c)$ is the pixel value of reconstructed image on channel c at position (i,j).

3.1. Comparison on different sources of fixation

The averages and standard deviations of RMSE and entropy on 100 reconstructed face images from FRGC data set [7] are compared on different sources of fixations. It can be seen that the averages and standard deviations of RMSE computed on actual fixations and model-generated fixations are lower than that computed on uniform and random distribution fixations. The table 1 demonstrates that the RMSE reconstructing with actual fixations and artificial model-generated fixations are much lower than that reconstructing with random, uniform fixations. It also shows that the average entropy of reconstructed images on different fixation source are almost equal.

3.2. RMSE and entropy with increasing fixations

We also find that the average RMSE in the face region between original images and reconstructed images decrease rapidly when the number of fixations increases. Result on Figure 1 indicates that we can reconstruct the face in ROI more accurate if we use more fixations, see Figure 4)

Meanwhile, the entropy in the face region almost holds constant after the number of fixations exceeds two (See Figure 5). It means that human capture most of information

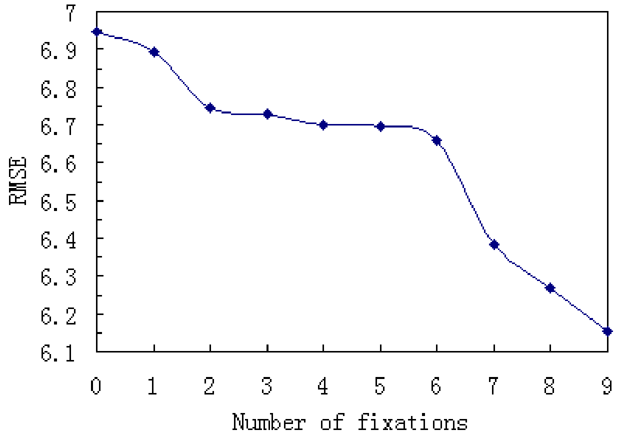

Figure 4. The RMSE decreases rapidly when reconstruct the face region with increasing number of fixations.

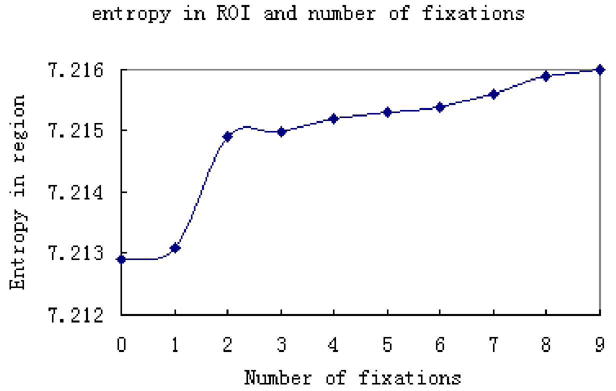

Figure 5. Entropy almost holds constant after the number of fixation reaches 2, the difference between maximum and minimum entropy is only 0.003.

with first two or three fixations, after that little information can be captured by subsequent fixations.

3.3. Statistical analysis of fixation distribution

As the work of Zhang et al. [18] indicated, facial feature extracted or selected usually distribute around key regions, such as eyes, nose and mouth regions. To verify if this phenomenon happens on actual eye fixations, we conduct a statistical experiment on which the distributions are computed using fixation positions data recorded on 4 subjects on 100 face images. The distributions are 2D histograms with 108 bins (12 rows and 9 columns), we linearly scale the value of histograms into range 0 to 255 and display them as images. (See Figure 7). It demonstrates that HVS actually puts more importance on regions like two eyes, nose and mouth regions.

The weights of different local regions on different scales (from [18]) are shown in Figure 6. We can see that two distributions are rather similar with each other, which suggest that face recognize systems can take the advantage weighted sampling of different regions in face image to achieve better recognition accuracy.

In our experiments, we find that people more likely to fixate at distinctive areas of face image such as an area with a mole on it. Figure 8 shows fixation distributions on a man with a mole on his chin and a woman with a adornment on the right corner of her eye.

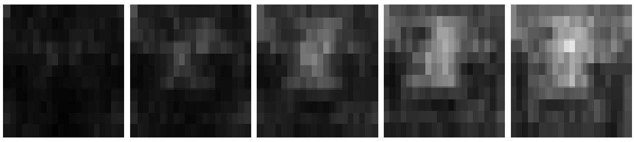

Figure 6. (Weights of different local regions for five scales learned from the FERET training set. [18](Courtesy of W. Zhang and et al.)

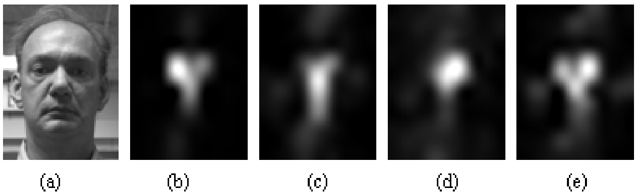

Figure 7. Fixation distribution statistics: (a) An sample image, (b), (c), (d), (e) Weights of different regions calculated on fixation positions of 4 subjects.

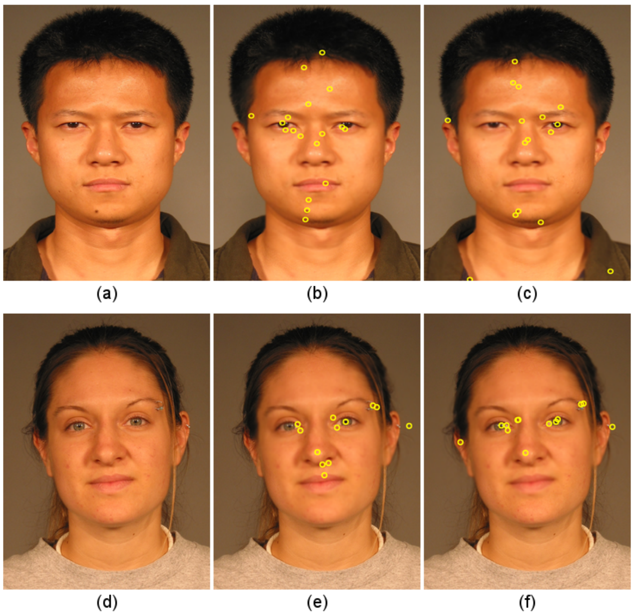

Figure 8. People tend to fixate on distinctive area of face. (Courtesy of FRGC [7] licenser).

4. CONCLUSION AND DISCUSSION

This paper proposes a simple yet flexible foveated imaging method for face reconstruction in ROI using actual, random, uniform and model-generated fixation positions and compare the effectiveness of different fixation source. Primary experiments on reconstruction of face region on 100 FRGC [7] face images using different fixation source have shown that the non-uniform sampling in HVS is optimal on capturing information contained in face image.

Foveated imaging method used in this experiment can been easily adjusted to better correspond with HVS by changing the parameters in this foveated imaging method. Future works on foveated imaging method include careful tuning of the parameters to better correspond with HVS and using it on other aspects such as image compression and real-time communications.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China under contract No.60673091 and No.60702031, Hi-Tech Research and Development Program of China under contract No.2006AA01Z122, Natural Science Foundation of Beijing under contract No.4072023, 100 Talents Program of CAS, and ISVISION Technology Co. Ltd.

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

[2] T. L. Arnow and A. C. Bovik. Foveated visual search for corners. *IEEE Transactions on Image Processing*, 16(3):813–823, 2007.

[3] S. Baluja and D. Pomerleau. Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Robotics and Autonomous System*, 22(3-4):329–344, Dec. 1997.

[4] W. S. Geiler and J. S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. *SPIE Proceeding*, 3299, 1998.

[5] W. S. Geisler and J. S. Perry. Real-time simulation of arbitrary visual fields. *ACM Symposium on Eye Tracking Research and Applications*, 2002.

[6] http://www.eyelinkinfo.com.

[7] http://www.frvt.org/FRGC/.

[8] http://www.intel.com/technology/computing/opencv/.

[9] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998.

[11] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[12] R. Milanese, S. Gil, and T. Pun. Attentive mechanism for dynamic and static scene analysis. *Optical Engineering*, 34(8):2428–2434, Aug. 1995.

[13] M. Mozer and S. S. *Attention*, chapter Computational modeling of spatial attention, pages 341–393. University College London, London, 1996.

[14] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L. N. Podladchikova, and N. A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Research*, 38:2387–2400, 1998.

[15] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *J. Electronic Imaging*, 10(1):152–160, 2001.

[16] P. Silsbee, A. C. Bovik, and D. Chen. Visual pattern image sequence coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 3:291–301, 1993.

[17] A. Yarbus. *Eye Movements and Vision*. New York: Plenum, 1967.

[18] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. *Proc. International Conference on Computer Vision 2005*, 1:786–791, 2005.