December 2003

Biologically Inspired Vision Sensor for the Detection of Higher-Level Image Features

Jan Van der Spiegel
University of Pennsylvania, jan@seas.upenn.edu

Masatoshi Nishimura
Sankyo Co., Ltd.

Recommended Citation

Jan Van der Spiegel and Masatoshi Nishimura, "Biologically Inspired Vision Sensor for the Detection of Higher-Level Image Features", . December 2003.

Copyright 2003 IEEE. Reprinted from *Proceedings of the 2003 IEEE Conference on Electron Devices and Solid-State Circuits*, pages 11-16.
Publisher URL: http://ieeexplore.ieee.org/xpl/tocresult.jsp?isNumber=28656

Biologically Inspired Vision Sensor for the Detection of Higher-Level Image Features

Abstract

The paper briefly reviews certain aspects of the biological visual system and presents a smart vision sensor for the detection of higher-level features. The visual system processes information in a hierarchical manner starting from the retina up to the visual cortex. It decomposes the image in simple features (edges, orientation, line stops, corners, etc) using spatial and temporal information. At the higher level it integrates these primitive features, resulting in the recognition of complex objects. The sensor described in the paper is loosely modeled after the visual system and incorporates pixel level, programmable elements which extract orientation, end stops, corners and junctions from a line drawing. The architecture resembles a CNN-UM that can be programmed with a 30-bit word. The 16 x 16 pixels array detects these higher-level features in about 54 µseconds.

Keywords

vision sensor, smart sensor, image features, biologically inspired, CNN

Comments

Biologically Inspired Vision Sensor for the Detection of Higher-Level Image Features

J. Van der Spiegel and M. Nishimura

Abstract - **The paper briefly reviews certain aspects of the biological visual system and presents a smart vision sensor for the detection of higher-level features. The visual system processes information in a hierarchical manner starting from the retina up to the visual cortex. It decomposes the image in simple features (edges, orientation, line stops, corners, etc) using spatial and temporal information. At the higher level it integrates these primitive features, resulting in the recognition of complex objects. The sensor described in the paper is loosely modeled after the visual system and incorporates pixel level, programmable elements which extract orientation, end stops, corners and junctions from a line drawing. The architecture resembles a CNN-UM that can be programmed with a 30-bit word. The 16x16 pixels array detects these higher-level features in about 54 μs.**

I. INTRODUCTION

Image sensors have benefited greatly from advances in CMOS technology, allowing increased performance and functionality at an ever-decreasing cost, faithfully following Moore's law. This has expanded considerably the application areas from traditional imaging to task-specific applications such as automotive, robotics, tracking, inspection, surveillance and security applications. These systems require real-time, low-cost and low-power solutions that can operate under a wide-range of illumination conditions. In many cases the goal is not to reproduce an image from the scene but rather extract *information* for further processing such as segmentation, classification and recognition. The traditional approach of capturing a high-quality image and subsequent processing by a PC or DSP is not always the optimal solution for these applications where real-time operation, size and power are main driving forces. Fig. 1(a) shows a conventional system in which the sensor and processor are physically separated through a high-bandwidth link. The advantage of this popular approach is the availability of high-resolution imagers and high performance processors (PCs or DSPs). However, image-processing algorithms are computationally intensive and require powerful processors. As a result, this solution may not be appropriate for applications where size and power are the key constraints. For such applications a more compact solution is desirable. The confluence of the emergence of Systems on a Chip (SoC) and an improved understanding of the biological visual system opens the possibilities for smart and efficient task-specific vision sensors. The approach is shown in Fig. 1 (b) where the imager performs both image capture and feature extraction functions. The on-chip processing consists usually of image pre-processing (compression, adaptation, contrast detection) and feature extraction (motion, position, orientation, etc.). These computations involve short-range interactions between neighboring pixel elements and are thus well suited for focal plane implementation. The end result is a smaller dataset with denser information content than the raw image. The added advantages are the relaxed bandwidth between the sensor and the CPU, and the need for a simpler processor (DSP or FPGA). Small and low-power realizations, such as a system in a package (SiP), now become a viable option.

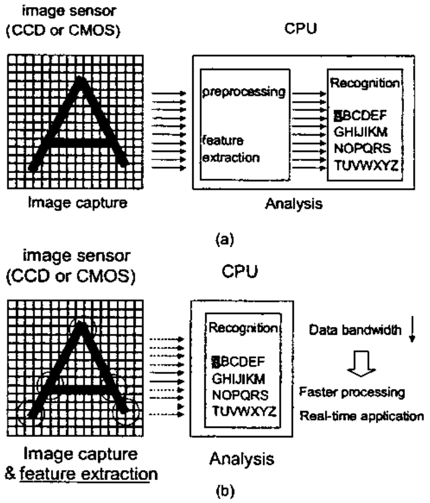


Fig. 1. Imaging system using (a) a conventional sensor, (b) a feature detection sensor as a front-end.

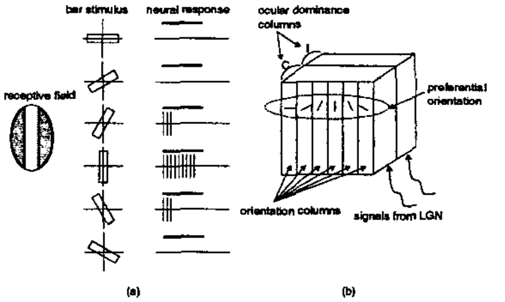
J. Van der Spiegel is with the Department of Electrical & Systems Engineering, at the University of Pennsylvania, Philadelphia, PA19107, USA, E-mail: jan@seas.upenn.edu. M. Nishimura is with Sankyo Co, Ltd, Tokyo 140-9710, Japan

0-7803-7749-4/03/$17.00©2003 IEEE.

II. Biological Vision System

The biological sensory system is amazingly efficient in image processing and pattern recognition in terms of speed, robustness and accuracy. It is highly structured and processes information in a hierarchical and parallel manner. The visual information processing starts at the retina where the light intensity is converted into electrical signals through cones and rods. In the outer plexiform layers of the retina the photoreceptors are connected to the horizontal and bipolar cells. The horizontal cells are mutually inhibited in a lateral direction producing a spatially smoothed version of the incoming signal. The bipolar cells receive excitatory input from the receptors and inhibitory inputs from the horizontal cells, resulting in an output that corresponds to the edges in the image. The edge detection is the result of the ON- and OFF-center receptive fields of the bipolar cells. A receptive field of a cell (neuron) is defined as the area of the retina that affects the cell's output. The receptive field is a key architectural strategy used at different levels of the processing chain. By changing the weights (value and sign) of the connections between the cells in the receptive field and the output neuron, various operations can be easily implemented. Edge detection in the retina is an efficient way to eliminate redundant information and highlight features of interest. Further modification is carried out in the inner plexiform layer where the ganglion cells integrate the contributions of the receptor, the horizontal and bipolar cells using the ON- and OFF-cells similarly as the bipolar cells. The output of the ganglion cells is a pulse train whose frequency is proportional of the signal strength. It is interesting to notice that the number of optic nerve fibers is 1 million whereas the number of photoreceptors is 100 million. This significant reduction is indicative of the pre-processing performed in the retina.

The ganglion cells transmit their signals through the lateral geniculate nucleus to the visual cortex. Area V1 of the cortex has three types of cells: simple, complex and hypercomplex cells. A simple cell responds to a bar of light with a specific orientation and position on the retina. Fig. 2 illustrates the cell's response when the direction of the bar is moved away from the preferred orientation. The orientation selectivity of the simple cell is obtained by combining the output of the ganglion cells aligned in a certain orientation. The complex cell also responds to a stimulus aligned in a certain direction but is position independent in contrast to the simple cell. The third type of cell is the hypercomplex cell and responds to more complex features of an image such as linestops and corners. The orientation detectors are arranged in columns in area V1, schematically shown in Fig. 2b. Each column is about 30-100 μm wide and 2 mm deep with a preferential orientation difference of 10° between neighboring columns. A set of columns that cover all orientations for each receptive field is called a hypercolumn [1]. Signals coming from the left and right eye are processed in closely spaced columns, called ocular dominance columns [2].



Fig, 2. Orientation selectivity of the cell response. (a) Response of the simple cell to stimuli of different orientations. (b) Hypercolumn structure.

Local information extracted from the retina is sent to neurons in the visual cortex that collect information from larger spatial regions, allowing the extraction of more complex and global features. Fig. 3 shows the shape of stimuli along the visual pathway from the retina (center-surround ON-center cell) through region V1 (simple, complex and hypercomplex cells) to the anterior part of the inferotemporal (AIT) cortex. The later one is believed to be responsible for object recognition. The complexity of the stimulus necessary for cell excitation as well as the size of the receptive field increases along this pathway. At the AIT for instance, cells that respond to T-type junctions, or star-shaped stimuli were reported [3]. At the final recognition layer complex objects are recognized through integration of the multiple representations of the image at each level of the hierarchy. In addition to the spatial aspects of the image, temporal features are equally important and are used for motion and tracking purposes.

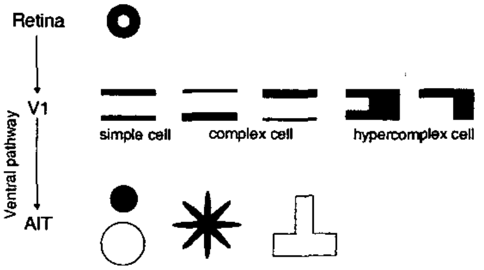


Fig. 3. Shape of the stimulus that excites the cells at several stages along the visual pathway.

In summary, a key computational strategy of the biological sensory system is the use of receptive fields to extract features and integrate these features in an hierarchical fashion to derive more complex ones [2]. These correspond to multiple representations along the different levels of the visual pathways, which are stored in local memory to aid with the computations of various features and improve visual performance [4]. This distributed architecture has the advantages of: (a) data reduction, (b) fast processing due to massive parallelism, and (c) robustness (to image deformations and illumination variability).

III. Neuromorphic Vision Sensors

The exceptional performance of the visual sensory information processing system in terms of sensitivity, robustness, signal-to-noise ratio, and recognition capabilities has provided strong incentives to build vision sensors modeled after the biological ones [5],[6],[7],[8] (and references therein). The focus of these papers has been either on improving the sensor performance or increasing the functionality. This has been obtained through inclusion of logarithmic response and adaptation techniques for increased dynamic range, the incorporation of processing elements for the detection of contrast and orientation [9-10], for motion estimation, texture classification [11] velocity detection [12-13], line orientation [14-15], and feature extraction for character recognition [16],[17],[18], among others. All of these sensors follow the approach of Fig. 1 (b) in which sensing and processing elements are integrated together into a smart pixel on the focal plane. In addition to incorporation of processing elements, the sampling structure of the retina also has been explored to improve functionality and performance. Space-variant sensors have been developed for tracking [19] and efficient image processing applications [20-21]. The spatial distribution can provide an additional dimension to improve the computational efficiency of the sensor [22].

IV. Sensor for Higher Level Features

Most of the papers published so far deal with image pre-processing and detection of low-level features. The remainder of this paper will focus on a smart sensor that extracts higher-level features (corners and junctions) to aid in the robust recognition of characters and objects. The motivation for such a sensor is the fact that corners and junctions carry rich information about the structure of an object and can help with object understanding and pattern recognition [23]. They are also important from a practical point of view since these are robust with respect to changes in perspective and small distortions. In addition, they are zero-dimensional and do not cause an aperture problem. Therefore they can be used to link multiple images from different perspectives [24]. The objective of the sensor is to derive and locate the following features: corners, T-, X-, Y-type junctions, and linestops. We will focus on line drawings which have been obtained by first extracting the edges of images or characters.

The sensor architecture is loosely based on the strategy employed in biology, which is described earlier in the paper. However, rather than implementing the feature extraction through multiple, parallel receptive fields [17], the sensor obtains a similar result through the use of programmable templates which are executed in sequence. A series of templates are applied in parallel over the whole image plane to extract various features. Thus, a trade-off between speed and area is made in order to minimize the total number of processing elements needed. Since silicon-based hardware is much faster than the biological wetware and can be reconfigured, this trade-off is easily justified and results in high-throughput feature extraction.

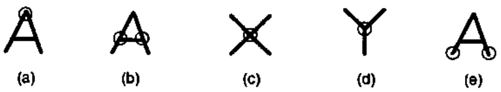


Fig. 4. Features of interest: corner, T-, X, Y-type junction and linestops.

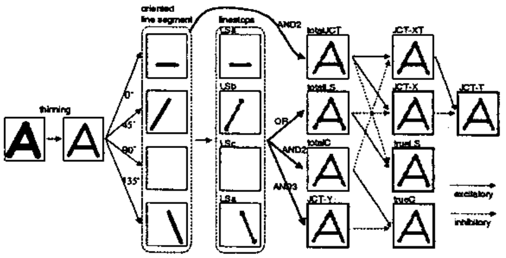


Fig. 5. Conceptual processing flow for the detection of the five features shown in Fig. 4.

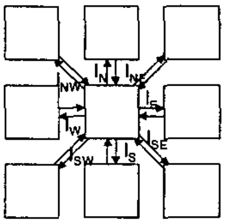
Fig. 4 shows the features of interest. It is clear that the five features are the result of interaction between line segments and linestops of different orientations. This leads to the conceptual processing flow given in Fig. 5. For a robust detection of the five features shown in the fat boxes on the right in Fig. 5, several intermediate processing steps (not shown in the figure) are required such as line completion, elimination of isolated points, line inhibition, line elongation, and thickening, in addition to the main steps of line thinning, orientation and linestop detection. These operations are performed independently for each of the four orientation planes (0,45, 90 and 135°). The operations can be carried out by template matching in a 3x3 window, in which the weights are specific for each type of operation,

$$x_{ij}(n+1)=f(\sum_{l=-1}^{1}\sum_{m=-1}^{1}x_{i+l,j+m}(n)\,r_{lm};I) \qquad (1)$$

in which $x_{ij}(n)$ is the binary status of the pixel at the position (i,j) at a discrete instant n, r_{ij} is the element of the template, f is the function to generate a binary output using the threshold I given in the form below:

$$f(x;I)=\begin{cases}1 & \text{for} \quad x\geq I\\ 0 & \text{for} \quad x<I.\end{cases} \qquad (2)$$

The template matching can be implemented as a convolution operation in which the kernel is the flipped version of the template in both horizontal and vertical directions. The convolution kernel, which represents the impulse response of a system, can be easily realized in hardware. A set of current outputs, each of which is proportional to the element specified in the convolution kernel, is generated from the pixel. By distributing these currents between neighboring pixels as shown in Fig. 6, the proposed algorithm can be mapped onto hardware. These distributed currents are summed at each pixel to produce the convolution result, which are then thresholded to generate a binary output voltage.



$$K=\begin{pmatrix}k_{-1-1} & k_{-10} & k_{-11}\\ k_{0-1} & k_{00} & k_{01}\\ k_{1-1} & k_{10} & k_{11}\end{pmatrix}=\begin{pmatrix}I_{NW} & I_{N} & I_{NW}\\ I_{W} & I_{C} & I_{E}\\ I_{WS} & I_{S} & k_{ES}\end{pmatrix}$$

$$k_{lm}=r_{-l-m}$$

Fig. 6. Implementation of a convolution kernel showing the current distribution.

A simplified pixel architecture is shown in Fig. 7 that consists of a photodetector, implemented as parasitic PNP transistor with floating base (nwell). The photocurrent is thresholded and stored in local memory when a control signal *photo* is activated. There are six memory registers of which four are used to store information about the four orientation planes and two as working memories. The processing stage distributes the reference current to neighbors according to a specified convolution kernel (not shown in the figure) and also sums the contributions from neighboring pixels. The sum is compared with the threshold I_{th2} in order to generate a binary output that is transferred into local memory. The actual signal processing is done in the analog domain while the control and storage is done in the digital domain. This allows optimal performance in terms of speed, accuracy and programmability. Extra care was taken in the design of the analog circuits to ensure that the error introduced by the current mirror circuits is small enough to prevent erroneous operation of the convolution while maintaining a fast enough response time [25]. The actual pixel contains additional switches to steer the currents according to the convolution kernel to the neighboring pixels. This makes the sensor programmable according to an external digital control. A 30-bit word is used to set the switches and determine the operation. As a result, the sensor is a simplified form of a cellular neural network universal machine (CNN-UM). The control words determine, the sequence of the various operations, including template matching and storing intermediate results.

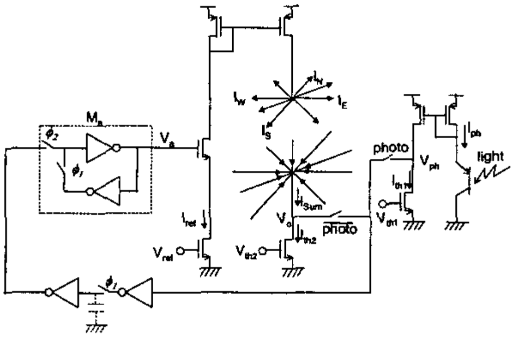


Fig. 7. Simplified pixel circuit.

The resulting pixel layout is shown in Fig. 8. The circuits are fabricated in a three metal, single poly 0.5 μm CMOS process. The pixel size is 154x153 μm^2 and fill factor is 12.5%. The phototransistor can be seen at the top of the pixel and the processing and memory elements are located underneath the photodetector. The signal lines are laid out to the right and bottom of the pixel, using the first two metal layers. The top metal layer acts as a light shield. A 16x16 experimental chip was fabricated to test the concepts and feature detection algorithms. The resulting chip size is 3.2x3.2 mm^2.

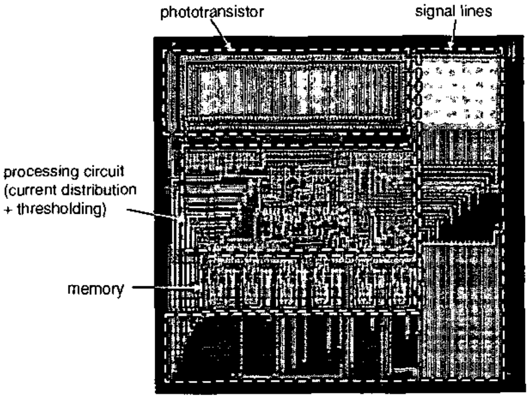


Fig. 8. Layout of a pixel.

V. Results

The sensor is tested by projecting various images on it. Measurements of the transistor mismatch shows that the value is well within the specified range resulting in a low error rate. Fig. 9 shows the response of the sensor to five letter images. The images shown in the bottom row are reconstructed by superimposing each feature at the detected position on the thinned image. All the important features are detected. Comparison of these results with the simulation result shows that the sensor response was identical to the one predicted by the simulation proving the validity of the approach. The sensor operates at a speed of 5MHz, which corresponds to 54 μs per feature detection involving over 270 individual processing operations.

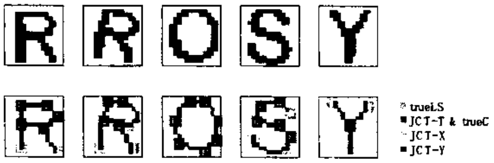


Fig. 9. Sensor response to various letters projected on the imager. The top row shows the original characters and the bottom row show the thinned letters and the corresponding final features detected.

VI. Discussion

Neuromorphic image sensors provide an alternative paradigm for a certain class of applications, which require real-time, low power, low cost and robust operation under wide variety of illumination conditions. Inherent to the sensor is the focal plane processing for edge detection, adaptation, motion and feature detection, among others. These operations involve pixel level processing which lends itself well for massively parallel focal plane implementation. Clearly, these sensors are not used for a faithful image rendering but for extraction of information as part of scene analysis or interpretation.

This paper describes a sensor that incorporates processing functions that is found in area V1 of the visual cortex. It detects higher-level features such as X-, Y- and T-type intersections as well as linestops. The sensor obtains this by a series of simple operations and combining the results of these operations to extract complex features, similarly as is done along the visual pathway. However, in contrast to the biological system the sensor detect these features using a sequence of programmable template matching operations. Intermediate results are stored in short-term memory to allow for further processing. Typical high-level feature detection requires around 270 individual operations and can be executed in about 54 μs. This relatively short time is possible thanks to the parallel, pixel level operations.

To illustrate the advantage of the focal plane, parallel processing, suppose that the same operations were performed on a processor running at 2GHz, and that a weighted sum at each pixel is computed in one clock cycle. For our 16x16 sensor the calculation for a single template matching operation over the entire array would take 1.15 μs. This results in a total of 31 ms for the 270 individual steps. This may not be a completely fair comparison since more efficient algorithms for serial machines can be used than the ones implemented on the sensors. However, it is fair to say that pre-processing of image data is efficiently done on the focal plain. The speed advantage of the proposed sensor becomes even more apparent for larger pixel arrays, since the advantage scales as N^2 in which NxN is the number of pixels of the array. The main drawback of the neuromorphic sensor is the reduced fill factor and lower resolution as a result of the added pixel complexity. This can be in part eliminated by using a more complex 3-D integration technology, as has been illustrated in [18]. However, for applications where high resolution is not the prime objective, neuromorphic sensors are a viable and cost effective solution.

Acknowledgement

The authors like to thank professor Paul Mueller for helpful suggestions.

References

[1] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106-154, 1962.

[2] G. M. Shepherd, *Neurobiology*, New York: Oxford Univ. Press, 1994.

[3] I. Fujita, K. Tanaka, M. Ito, and K. Cheng, "Columns for visual features of objects in monkey inferotemporal cortex," *Nature*, vol. 360, pp. 343-346, 1992.

[4] B. Wandell, A. El Gamal and B. Girod, "Common Principles of Image Acquisition Systems and Biological Vision," *Proc. IEEE*, vol. 90, pp. 5-17, 2002.

[5] C. Koch and H. Li (eds), *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, Piscataway, IEEE Computer Press, 1995.

[6] C. Mead, *Analog VLSI and Neural Systems*, Reading: Addison Wesley Publ, 1989.

[7] A. Moini, *Vision Chips*, Norwell: Kluwer Academic Publ, 1999.

[8] J. Van der Spiegel and R. Etienne-Cummings, "Neuromorphic Vision Sensors," *Sensors and Actuators A*, vol. 56, pp. 19-29, 1996.

[9] P-F. Ruedi et al, A 128x128 Pixel 120dB Dynamic Range Vision Sensor Chip for Image Contrast and Orientation Extraction," *Digest IEEE Int. Solid-State Circuits Conf.*,vol. XLVI, pp. 226-227, 2003.

[10] M. Barbaro et al., "A 100x100 Pixel Silicon Retina for Gradient Extraction with Steering Filter Capabilities and Temporal Output Coding," *IEEE J. Solid-State Circuits*, vol. 37, pp.160-172, 2002.

[11] R. Dominguez-Castro et al., "A 0.8mm CMOS Two-Dimensional Programmable Mixed-Signal Focal-Plane Array Processor with On-Chip Binary Image and Instruction Storage," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1013-1026, 1997.

[12] H.-C. Jiang and C.-Y. Wu, "A 2-D Velocity- and Direction-Selective Sensor with BJT-Based Silicon Retina and Temporal Zero-Crossing Detector," *IEEE J. Solid-State Circuits*, vol. 34, pp. 241-247, 1999.

[13] R. Etienne-Cummings, J. Van der Spiegel and P. Mueller, "A Focal Plane Visual Motion Measurement Sensor," *IEEE Trans. Circuits and Systems*, Part I, vol. 44, pp.55-66, 1997.

[14] R. Etienne-Cumming, Z. Kalayjian and D. Cai, "A Programmable Focal-Plane MIMD Image Processor Chip, *IEEE J. Solid-State Circuits*, vol. 36, pp. 64-73, 2001.

[15] P. Venier, A. Mortara, X. Arrogate and E. Vittoz, "An Integrated Cortical Layer for Orientation Enhancement," *IEEE J. Solid-State Circuits*, vol. 32, pp. 177-186, 1997.

[16] B. M. Bo, D. Caviglia and M. Valle, "An Analog VLSI Implementation of a Feature Extractor for Real Time Optical Character Recognition," *IEEE J. Solid-State Circuits*, vol. 33, pp. 556-563, 1998.

[17] W. Camp and J. Van der Spiegel, :"A Silicon VLSI Optical Sensor for Pattern Recognition," *Sensors and Actuators A*, vol. 43, pp. 188-195, 1994.

[18] M. Koyanagi, Y. Nakagawa, K.W. Lee et al, "Neuromorphic Vision Chip Fabricated using Three-Dimensional Integration Technology," *Digest IEEE Int. Solid State Circuits Conf*, vol. XLIV, pp. 270-271, 2001.

[19] R. Etienne-Cumming, J. Van der Spiegel, et al., "A Foveated Silicon Retina for Two-Dimensional Tracking," *IEEE Trans. Circuits and Systems II*, vol. 47, pp. 504-517, 2000.

[20] "A Foveated Retina-Like Sensor Based on CCD Technology," *Analog VLSI Implementation of Neural Systems"*, Chapter 8, pp. 189-210, eds. C. Mead and M. Ismail, Boston: Kluwer Academic Publ., MA, 1989.

[21] F. Pardo, B. Dierickx, D. Scheffer, "Space-Variant Nonorthogonal Structure CMOS Image Sensor Design," *IEEE J. Solid-State Circuits*, vol. 33, pp. 842-849, 1998.

[22] G. Kreider, "A Treatise on Log-Polar Imaging Using a Custom Computational Sensor," Ph.D. Dissertation, Dept. Electr. Eng., Univ. of Pennsylvania, Philadelphia, PA 1993.

[23] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183-193, 1954.

[24] B. S. Manjunath, C. Shekhar, and R. Chellappa, "A new approach to image feature detection with applications," *Pattern Recognition*, vol. 29, pp. 627-640, 1996.

[25] M. Nishimura, "A VLSI Computational Sensor for the Detection of Image Features," Ph.D. Dissertation, Dept. Electr. Eng., Univ. of Pennsylvania, Philadelphia, PA, 2001.