Submitted to DARPA Image Understanding Workshop 1997
May 11–14, 1997, New Orleans, Louisiana, USA

# Sensory Attention: Computational Sensor Paradigm for Low–Latency Adaptive Vision

**Vladimir Brajovic and Takeo Kanade**
School of Computer Science, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh PA 15213
brajovic@cs.cmu.edu, tk@cs.cmu.edu
http://www.cs.cmu.edu/afs/cs/usr/brajovic/www/lab/vlsi.html

## Abstract[1]

The need for robust self–contained and low-latency vision systems is growing: high speed visual servoing and vision–based human computer interface. Conventional vision systems can hardly meet this need because 1) the latency is incurred in a data transfer and computational bottlenecks, and 2) there is no top–down feedback to adapt sensor performance for improved robustness. In this paper we present a tracking computational sensor — a VLSI implementation of a sensory attention. The tracking sensor focuses attention on a salient feature in its receptive field and maintains this attention in the world coordinates. Using both low –latency massive parallel processing and top–down sensory adaptation, the sensor reliably tracks features of interest while it suppresses other irrelevant features that may interfere with the task at hand.

## 1. Introduction

The computational sensor paradigm [Kanade and Bajcsy, 1993] has the potential to greatly reduce latency and provide top–down sensory adaptation to vision systems. By integrating sensing and processing on a VLSI chip, both transfer and computational bottlenecks can be alleviated; on–chip routing provides high throughput transfer, while an on-chip processor could implement massively–parallel fine–grain computation, thus providing high processing capacity which readily scales up with the image size. In addition, the tight coupling between processor and sensor allows for efficient top–down feedback that can control and adjust sensor for further acquisition based on the preliminary results of the processing. Our recent work has been concerned with efficient implementation of global operations over a large group of image data using the computational sensor paradigm [Brajovic and Kanade, 1994]. We have formulated two mechanisms for implementing global operations in computational sensors: (1) *intensity–to–time processing paradigm* [Brajovic and Kanade, 1996], and (2) *sensory attention* presented in this paper.

## 2. Approach

The sensory attention is based on the premise that salient features within the retinal image represent important global features of the entire image. By selecting a small region of interest around the salient feature for subsequent processing, the sensory attention eliminates extraneous information and allows the processor to handle small amounts of data at a time. We have implemented sensory attention by fabricating and testing *tracking computational sensor.* The tracking computational sensor optically receives a saliency map and continuously selects and tracks the peaks in it. The location and intensity of the selected peaks is reported on few output pins with low latency. These quantities are also used internally in a top–down fashion to aid tracking of the attended location. The chip is a 28 x 28 array of 60μ x 60μ cells, and is fabricated on a 2.2mm x 2.2mm die.

The *sensory attention* follows the model of *visual attention* in brains. This analogy is attractive for

1. This research has been sponsored by Office of Naval research (ONR) under Contract N00014–95–1–0591. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ONR or the U.S. Government.

two reasons. First, the main argument that has been used to explain the need for selective visual attention in brains is that there exist some kind of processing and communication limitation in the visual system. So it does in machines. Attention "funnels" only relevant information and protects the limited communication and processing resources from the information overload. Second, it has been shown that the visual attention improves performance, and is needed for maintaining coherent behavior while interacting with the environment (i.e., attention–for–action) [Allport, 1989]. Unlike eye movement (i.e., *overt* shifts), the attention shifts (i.e., *covert* shifts) do not require any motor action, but occur internally on a fixed retinal image. For this reason, attention shifts are faster and play an important role in low–latency vision systems.

It is interesting to note that foveating computational sensors [Kanade and Bajcsy, 1993] try to emulate this kind of data compression. For example, Van der Spiegel's log–polar sensor samples images within fovea with high acuity, while maintaining sparse representation at the periphery. This sensor simulates overt shifts, since it requires motor action for foveating. Kosonocky's foveating sensor allows programmable fovea within the retinal image; therefore, it eliminates the need for mechanical action and simulates covert shifts. Another related solution is random access to the image data. For example, Laval's MAR sensor attends to, and reads only, a small local portion of the retinal image, the part that is necessary for the local convolution performed in the global off–chip processor. However, these computational sensors act as special cameras and the mechanism which guides the location of the attention is missing.

To apply attention selection in machines, several issues must be solved: (1) the problem of selecting an "interesting" location, (2) the problem of shifting to another location, and (3) the problem of transferring local data for further processing. In a very influential paper [Koch and Ullman, 1987], Koch and Ullman address these issues. The selection process utilizes a *saliency map* that encodes conspicuousness or the level of interest throughout the retinal image. The saliency map can be derived from image features, including: intensity, color, spatial and temporal derivatives, motion, and orientation. For selecting a location of the attention within the saliency map, *winner–take–all* (WTA) mechanism has been suggested. The WTA is not responsible for information processing; rather it determines only which area of the retinal image should be relayed to the global processor for further inspection.

The problem of shifting to another location is somewhat more challenging. It is observed in humans that interesting visual stimulation initially (i.e., during the first 100ms) captures the attention; later (i.e., after 300ms) it has inhibitory effects which can last up to 1.5 seconds [Milanese, 1993]. The inhibitory effect prevents the subject from returning to previously visited locations. The inhibition is "stored" in environmental coordinates rather than in image coordinates; therefore, reliable operation is maintained even in the presence of ocular or object movement. The attention shifts can be initiated on a voluntary basis by telling the observer the location of a target, or they can be automatic caused by the onset of a visual stimulus. For shifting to another location, Koch and Ullman's model allows the saliency of the currently attended location to decay, even if the visual stimuli creating the saliency remain present. This will release the WTA mechanism and allow it to converge to another location. Either a *local* or *central* inhibition mechanism for initiating decay is possible. The local mechanism causes the saliency to decay some time after the WTA has converged to a particular location. In the central mechanism, once the attended portion of the retinal image is relayed to the central processor, a signal, which inhibits the conspicuousness of the currently attended location, is sent back. The local inhibition mechanism mimics the automatic attention shift, while the central mechanism can initiate voluntary attention shifts.

Recently, Morris et al. [Morris and DeWeerth, 1996] reported an analog VLSI circuit implementation of covert attention shifts as suggested by the Koch and Ullman model. A one–dimensional 19 cell circuit implements: 1)saliency map normalization, 2)WTA location selection with preference for spatial proximity shifts, 3)inhibition of return control and 4) position detection for producing the location of attention as the output. Depending on the biasing condition, the circuit is able to roam between the peaks in the stationary saliency map.

In *the attention–for–action* model, Allport sug-

gested that attention goes beyond protecting the limited processing resources during complex object recognition: *attention is needed to ensure behavioral coherence* [Allport, 1989]. Since visual perception is the means for allowing a subject to interact with the environment (e.g., manipulate, avoid, etc.), it must produce actions consistent with the subject's goals. Selective processing is necessary in order to isolate the information that defines parameters for the appropriate action. For example, to catch a moving object, among many other moving and stationary objects, the information specific only to that object determines the action. Information about other objects in the visual field must be kept from interfering with the goal of catching the target object, even though other objects may influence how the target object is caught. In other words, attention aids the target goal by masking the irrelevant information's interference, but allows the action to be modified or diverted if new, important events occur.

The attention–for–action model is in close agreement with our goal of producing reliable *low–latency* computational sensors which provide *useful* information for the *coherent interaction with the environment*. It is not hard to imagine that if the attention is allowed to arbitrarily roam from one location to another, as suggested by Koch and Ullman's model and implemented in [Morris and DeWeerth, 1996], it may take a long time before the global processor encounters the *relevant* information for an appropriate action. We need more control over attention shifts, possibly by employing the *central inhibition* mechanism in combination with the *voluntary focus of attention* directed toward desired goals. For robust operation, such shifts must maintain the location of attention in the presence of ocular or object motion [Milanese, 1993].

## 3. Implementation

In the prototype implementation of the sensory attention proposed by this work, our concern is not how to compute the saliency map, but rather how to quickly and reliably locate and maintain an interesting location in the saliency map. We call this embodiment of the sensory attention *tracking computational sensor* because, when the saliency map is a natural image — the trivial saliency map — the features that attract attention are bright spots in the environment. The tracking computational sensor selects and tracks those spots while ignoring the background.

### 3.1. Location Selection

An image representing a saliency map is focused onto the array of photo detectors: photodiodes or photo transistors. The generated photo currents are fed to the winner–take–all (WTA) circuit which is responsible for the detection of the maximum point. The selected location is called a *feature*. Our design is based on a WTA circuit originally proposed in [Andreou et al. 1992] and [Lazzaro et al., 1988] shown is Figure 1. Currents $I_1 \ldots I_N$ are the input photo currents, while currents $J_1 \ldots J_N$ are the outputs of the WTA circuit. The cell receiving the largest photo current $I_k = max(I_1 \ldots I_N)$ responds with non–zero output current $J_k = I_c \neq 0$, while other cells respond with zero currents, i.e., $J_i = 0$, for $i \neq k$. The peak photo current establishes and holds the common voltage $V_c$. For small input currents, like those produced by light detection, the transistor operates in the subthreshold region. In that case, the voltage $V_c$ is the logarithm of the winning input current: $V_c = V_o \log(I_1 / I_o)$, where $I_o$ is the process parameter and $V_o = kT/q\kappa$. Therefore, the intensity of the winner is accessed globally by monitoring the voltage on the common wire.

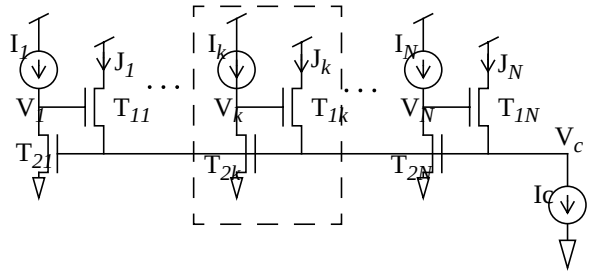


**Figure 1:** Schematic diagram of the winner–take–all circuit. Boxed area indicates one cell.

Since only the winning cell responds with non–zero current, the WTA effectively provides 1–of–N binary encoding of the winner's position. A digital on-chip decoder easily converts this code to any other binary code such as a natural binary or BCD code. In addition, there are efficient analog means for winner localization [DeWeerth, 1992]. In one example, the outputs from each WTA cell are con-

nected to nodes of a linear resistive network. The network behaves as a current divider splitting current $I_c$ into two peripheral components, each proportional to the position of the current injection. By reading these peripheral components, the location of the winning cell is found. The WTA cells can be physically laid out in a two–dimensional array. Using the method of projections [Horn, 1986], the position of this current in two dimensions is found by solving two one–dimensional problems. Two copies of the output current are summed into the horizontal and vertical bus, respectively. The total current in these buses represents the desired projections onto the *x* and *y* axes. Then, two linear resistive networks are used at the periphery of the array to locate the winner in a *x* and *y* direction.

### 3.2. Location Shifts

The two dimensional WTA circuit locates the absolute maximum in the entire saliency map. In practical applications, there are often several strong features in the saliency map which are candidates for attracting the attention. For implementation of the attention–for–action model, we need to direct attention toward a feature that is useful for the task at hand. This corresponds to voluntary attention shifts, i.e., "telling" the sensor where to "look." Once the feature is selected, we need a mechanism that will track the feature and thus maintain the location of attention in the environmental coordinates even in the presence of ocular motion. Our implementation inhibits portions of the saliency map and restricts the activity of the WTA circuit within a programmable active region within the whole array of photo receptors. The active region is programmed by appropriate row and column addressing, and corresponds to the central inhibition control suggested by Koch and Ullman.

There are two modes of operation: (1) select mode, and (2) track mode. In the *select mode*, the active region is defined by the external addressing (Figure 2a). The active region can be of arbitrary size and location. The sensor selects the absolute maximum within this region. In the *tracking mode* the sensor itself dynamically defines a small (e.g., 3 x 3 in our implementation) active region centered around the most recent location of the attention (Figure 2b).

The *select* mode directs the attention towards a feature that is useful for the task at hand. For example, a user may want to specify an initial active region, aiding the sensor to attend to a relevant local peak in the saliency map. Then, the *tracking* mode is enabled for locking onto the selected feature. The ability of the sensor to define its own active region is an example of the top–down sensory adaptation presently missing from conventional vision systems.

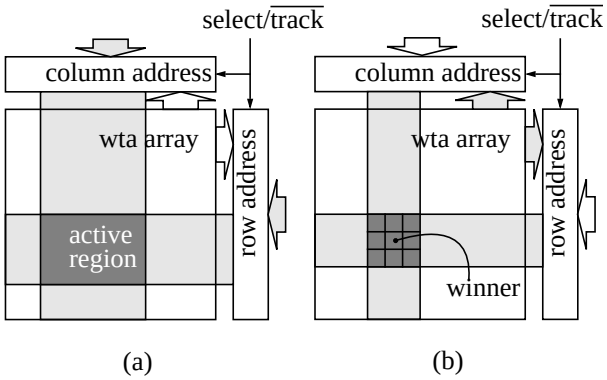


**Figure 2:** Modes of operation for the sensory attention computational sensor: (a) select mode, and (b) tracking mode.

The active region is programmed by inhibiting particular WTA cells under the external control. A circuit diagram of the WTA cell with inhibition is shown in Figure 3. The shunting path for the photo current is provided through the transistors $T_5$ and $T_6$. To maintain the cell active both $\overline{col}$ and $\overline{row}$ signals must be asserted (i.e. must be zero).

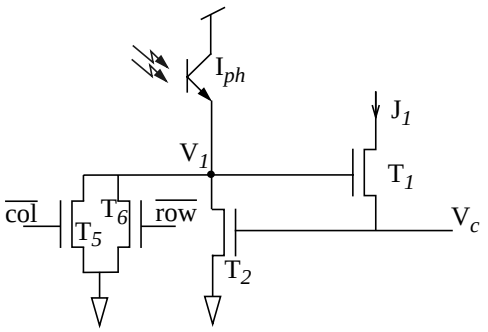


**Figure 3:** WTA cell with inhibition. (Shadowed area indicates components for cell inhibition.)

The control of active region is achieved from the periphery of the two–dimensional WTA array. The peripheral logic across three columns is shown in Figure 4. Similar logic is implemented for row addressing. In the select mode, the active column band is programmed by the content of the shift reg-

ister. There are no restrictions on the width or location of the band, as any bit pattern can be entered into the shift register.

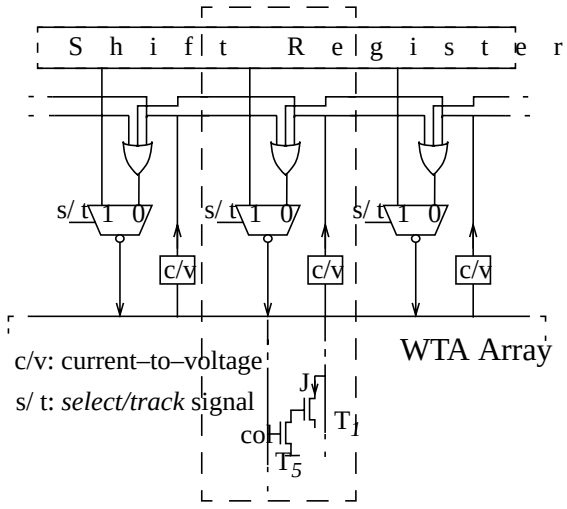


**Figure 4:** Peripheral logic for central control of the active region. The boxed area indicates one column. Similar logic is used for row addressing (not shown).

In the track mode, the active region is programmed by the WTA array and is dependent on the location of the feature being tracked. A particular column is enabled if the winning feature is on that column, or on one of the two immediate neighbors. In conjunction with the row inhibition (not shown), the tracking mode programs a 3 x 3 active region centered on the most recent feature. If that feature starts moving, one of the eight active neighbors will receive the winning feature and automatically update the position of the 3x3 active region. It is now clear that the salient feature is not necessarily the absolute maximum in the field–of–view, but rather it is a local peak in the retinal image. If for any reason the tracking mode starts on a location which is not a local peak, the 3x3 active region will "slide" along the intensity gradient until it locks onto a nearby peak.

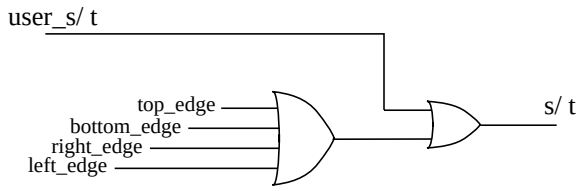


**Figure 5:** Logic for automatic switching between *select* and *track* modes.

With moving objects, the feature which is being tracked may reach the sensor's edge and fall out of the field of view. In order to ensure coherent transition in these situations, the logic shown in Figure 5 is implemented. The user may define the select mode by asserting signal *user_s/t*. However, when the user enables the tracking mode, the active region will be of size 3 by 3 as long as the tracked feature is not on one of the four edges of the array. When the feature reaches one of the four edges, the sensor automatically goes to a select mode. For a moment, the active region specified in the shift registers is enabled, and the absolute maximum is selected therein. If the newly selected feature is no longer on the edge, the sensor automatically goes back to the tracking mode, shrinks the active region to a 3 by 3 size, and continues feature tracking.

### 3.3. Transferring Local Data

Once the relevant conspicuous point has been localized in the saliency map, the local data from the attended vicinity must be transferred to the global processor for decision making. The local data originate from any early representation including: image data, early features used for building the saliency map, or the saliency map itself. The circuit for sensory attention described so far only receives and has access to the saliency map. However, with the suggested implementation, the local information from the saliency map can be easily transferred to the global processor. In fact, the magnitude of the localized feature in the saliency map is continuously reported to the global processor, as it is inherently measured by the WTA circuit. If the surrounding points are also needed, the global processor can program a trivial 1 x 1 active region at the desired location. The global processor inhibits all inputs of the saliency map except the programmed cell, and forces the WTA circuit to choose that particular point as the winner and report its magnitude on the global wire. We scanned the 1 x 1 active region throughout the array and collected several images (Figure 6).


**Figure 6:** Images from the tracking sensor (24x24 pixels).

## 4. Experimental Evaluation

Two tracking sensors prototypes — 1D and 2D —

have been built and tested for static and dynamic performance. The static performance has been tested on an early 1D prototype with 20 cells fabricated in 2μ CMOS technology. The findings have been reported earlier in [Brajovic and Kanade, 1994].

The temporal response of the WTA circuit is important when tracking moving features within dynamic saliency map. The dynamics of the circuit is a function of the parasitic capacitance at the input node *V1* comprising capacitance of the photo detector and capacitances of the gates and drains attached to this node. For a cell to win or lose this capacitance must be charged and discharged with the photocurrent. For average room illumination the photo currents are very small, much less then 1nA. Therefore, the WTA circuit in its original configuration is slow. To improve the dynamic performance of the WTA circuit, several measures can be taken: (1) increase the photo current, (2) decrease parasitic capacitance C, and (3) reduce the voltage swing on the capacitance *C*. A modified WTA cell that implements all three of these measures is shown in Figure 7. The photo transistor amplifies the photo current, $T_3$ isolates capacitance of the photo detector, and $T_4$ acts as a pull–up and limits the voltage swing.

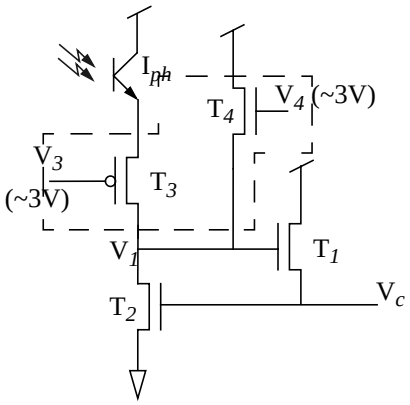


**Figure 7:** WTA cell with improved dynamic performance. (Fenced area indicates additions to the original WTA cell.)

The dynamic performance is evaluated for a 28x28–cell two–dimensional tracking computational sensor. Each cell is 62μ square. The photo transistor occupies about 30% of the cell's area. In the experimental set up, a scanning mirror reflects a beam of light onto a white cardboard. This produces a dot which travels along a straight line. The tracking sensor images the scene and tracks the moving dot. The rows of the sensor are approximately aligned with the trajectory of the laser dot, so that only *x* position needs to be observed. The mirror is driven from a sinusoidal oscillator whose frequency is adjustable. The maximum instantaneous velocity is attained at the middle of the trajectory. The goal is to observe how quickly the tracking sensor can shift attention, that is, how quickly it can update the feature's location as the feature travels across the array of cells. From the geometry of the set up, we can derive feature velocity from the frequency of the scanning mirror and then express it in image coordinates.

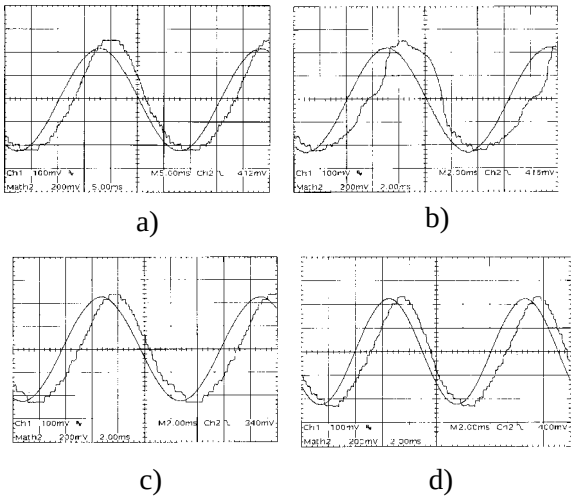


**Figure 8:** Tracking performance a) without the current buffer and without the pull–up. f=33Hz., b) without the current buffer and without the pull–up. f=83Hz., c) Tracking performance with the current buffer but without the pull–up. f=83Hz., d) Tracking performance with both the current buffer and

The effects of the current buffer and the pull–up can be turned on or off by biasing $V_3$ and $V_4$. Without the buffer and the pull–up, the sensor was reliably tracking up to the scanning frequency of 33Hz or 2,303.6 cells/second. Figure 8a shows two measured waveforms: (1) the feature's position *x* as reported by the tracking sensor, and (2) the sinusoid driving the mirror. If the frequency of the mirror is further increased, the reported position begins to distort. This is illustrated in Figure 8b for the scanning frequency of 83Hz. The tracking capability of the sensor starts to break down in the middle of the trajectory, as the velocity of the feature is the greatest there. Then, the current buffer is turned on by biasing $V_3$. The dynamic performance improved: the maximum tracking frequency is

increased from 33Hz to about 83.3Hz or from 2303.6 to 5793.9 cells/second. This is shown in Figure 8c; previously distorted waveform for the feature's position now better resembles the sinusoid. Finally, the pull up transistor is turned on by biasing $V_4$. The dynamic performance is slightly improved as shown in Figure 8d — the feature tracking is improved from 83Hz to about 100Hz, or to 6980.6 cells/second.

Another set of experiments is performed to evaluate how the intensity of the feature influences the dynamic performance. Using neutral density filters placed in front of the sensor's lens, the light is controllably attenuated. For each filter, the frequency of the mirror is increased until the waveform of the feature's position begins to distort. In this way, the maximum frequency is estimated for each intensity. Two sets of experiments are performed: (1) without the buffer and the pull–up, and (2) with the buffer and the pull–up. The results are graphed in Figure 9.

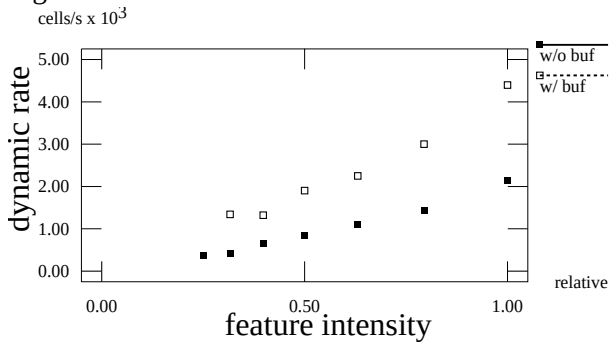


**Figure 9:** Maximum angular velocity of the attention shifts as a function of the relative feature intensity.

## 5. Conclusion

The proposed implementation for the sensory attention exhibits several interesting features. It performs a global operation over the saliency map and produces few global results: the position and magnitude of the selected saliency feature. These global results can be routed off–chip with low latency via few output pins. Furthermore, in the tracking mode, the global results are used internally for programming a 3 x 3 active region. This a top–down feedback secured robust performance in tracking the feature of interest while ignoring interference from other potentially stronger sources.

## References


[Allport, 1989] Allport, A. "Visual Attention," Foundation of Cognitive Science, M. Posner (ed.), MIT Press, 1989, pp. 631–682.

[Andreou et al. 1992] A.G. Andreou, et al., "Current-Mode Subthreshold MOS Circuits for Analog VLSI Neural Systems," *IEEE Trans. on NN*, Vol. 2, No. 2, pp. 205-213, March 1992

[Brajovic and Kanade, 1994] Brajovic, V. and T. Kanade, "Computational Sensors for Global Operations," *IUS Proceedings*, pp. 621-630, 1994.

[Brajovic and Kanade, 1996] Brajovic, V. and T. Kanade, "A Sorting Image Sensor: An Example of Massively Parallel Intensity-to-Time Processing for Low-Latency Computational Sensors," Proc. of the 1996 IEEE Intl. Conf. on Robotics and Automation, April 1996, Minneapolis, MN.

[DeWeerth, 1992] DeWeerth, S.P., "Analog VLSI Circuits for Stimulus Localization and Centroid Computation," *Intl. Jour. of Comp. Vision*, Vol. 8, No. 3, 1992, pp. 191-202.

[Horn, 1986] Horn, B., *Robot Vision*, MIT Press, 1986.

[Kanade and Bajcsy, 1993] Kanade, T. and R. Bajcsy, "Computational Sensors: A Report from DARPA workshop", *IUS Proceedings*, 1993.

[Koch and Ullman, 1987] Koch, C. and S. Ullman, "Shifts in Selective Visual Attention: Toward the Underlying Neural Circuitry. In L.M. Vaina (edt.), *Matters of Intelligence*, Reidel Publishing, 1987, pp. 115-141.

[Lazzaro et al., 1988] J. Lazzaro, S. Ryckebusch, M.A. Mahowald and C. Mead, "Winner-Take-All Networks of O(n) Complexity," in *Adv. in Neural Inf. Proc. Sys. Vol. 1*, D. Tourestzky, ed., pp. 703-711, Morgan Kaufmann, San Mateo, CA, 1988.

[Milanese, 1993] R. Milanese, "*Detecting Salient regions in an Image: From Biological Evidence to Computer Implementation,*" Ph.D., Dept. of Com. Sci., U. of Genova, Switzerland, Dec. 1993.

[Morris and DeWeerth, 1996] T.G. Morris and S.P. DeWeerth, *Analog VLSI Circuits for Covert Attentional Shifts*, MicroNeuro 1996, Lausanne, Switzerland.