Carnegie Mellon University
Doctoral Thesis in Robotics

Computational Sensor for Global Operations in Vision (Summary)

Vladimir Brajovic

January 22, 1996

Perception plays a dominant role in the development of an intelligent behavior. Our awareness of the environment relies on the activity of our sense organs. These outposts of the nervous system translate environmental changes into activity in sensory nerve fibers. It is then the function of the central nervous system to interpret this sensory information, integrating it into an appropriate pattern of behavior. Like biological systems, intelligent robotic behavior relies heavily on the sensory perception. Especially rich in information, and fascinating in its capability, is *vision*. It is not surprising that vision research has received equally high interest in neurophysiology, psychology, computer science and engineering.

In the last 30 years machine vision research advanced along many fronts. Cameras have improved: their resolution and sensitivity have increased, and new sensors such as uncooled infrared cameras are now commercially available. Many recognition algorithms have been developed: from 3D model matching to artificial neural networks. Yet performance of the existing machine vision systems still significantly lags that of biological vision. The two most critical features presently missing from the machine vision are *low latency processing* and *top–down sensory adaptation*.

The main contribution of this thesis is towards overcoming these two deficiencies by *implementing global operations in computational sensors*. Additional aims are to produce task–oriented self–contained machine vision components that can be used by a machine for a coherent interaction with the environment.

1 Motivation

The fundamental problem in machine vision comes from the computational complexity of basic tasks. Examples include the problem of detecting a target element in an image (*visual search*) and the problem of finding a correspondence between the image and a set of models (*matching*). Any algorithm which solves these problems in a general way, without the help of assumptions and heuristics, requires exponential execution time as a function of the image size and the number of stored models. From this observation it becomes apparent that vision systems have limited capability to scale up with images of increasing size and complexity.

The consistent paradigm in machine vision has been that a "camera" sees the world and a computer "algorithm" recognizes the object. Implicit in this view is the separation between the camera — a sensing device for transducing spatio–spectral–temporal phenomena to electric signals, and the computer — a computational device for processing and make sense out of data. That is, the transduced signal is read out of the sensor and digitized into the computer for processing. The separation of sensing and processing has resulted in several deficiencies in the computer vision systems developed so far. The two most critical features missing from the sequential paradigm are *low latency processing* and *top–down sensory adaptation*.

Latency, or reaction time, is the time that a system takes to react to an event. For example, a standard video camera takes 1/30 of a second to transfer an image. In many robotic applications it is too late by the time the system receives the image from such a camera. As another example, pipelined dedicated vision hardware can deliver the processing power to update its output 30 times per second, but the latency incurred through the pipeline is typically several seconds. These examples point to two primary sources of latency in vision systems: *the data transfer bottleneck* caused by the need to transfer an image from the camera to the processor, and *the computational load bottleneck* caused by the processor's inability to quickly handle the large amount of data. The detrimental effects of both bottlenecks scale–up with the image size.

Another aspect that has been neglected in machine vision is the top–down sensory adaptation. Many learning algorithms have been developed that adjust to variations in appearance of an object in sensor images. Nevertheless, complex ad–hoc algorithms that try to extract relevant information from inadequate sensor data are inevitably unreliable. In fact, time and time again it has been observed that using the most appropriate sensing modality

or setup, allows recognition algorithms to be far simpler and more reliable. For example, the concept of active vision proposes to control the geometric parameters of the camera (e.g. pan, tilt, etc.) to improve the reliability of the perception [4]. It has been shown that initially ill–posed problems can be solved after the top–down adaptation of the camera's pose has acquired new more appropriate image data. However, adjusting geometric parameters is only one level at which adaptation can take place. A system that can adjust its operation at all levels, even down to the point of sensing, would be far more adaptive than the one that tries to cope with the variations at the "algorithmic" or "motoric" level alone.

The lack of fast processing and top–down sensory adaptation in the sense–then–process paradigm, suggest that an alternative is needed.

Compared to the capabilities of the available machine vision systems and techniques, the performance of biological vision is astonishing. It has been estimated that humans can recognize up to 100,000 objects within 100–200 ms [57]. In addition, the recognition has a high degree of invariance with respect to factors such as the position, scale and orientation, which may completely change the retinal image of objects.

One of the most important factors which determines these capabilities is the high number of processing elements (approx. 10^{11} neurons) working in parallel in the human brain. However, given the relatively slow response of each neuron and the huge amount of input data (approx. 10^8 receptors), it becomes apparent that the sheer number of neurons is not sufficient to explain these performances. In fact, the human visual system is not even structured to exploit the computational power of a single, fully–connected network of cells; it is rather organized into a number of areas analyzing different aspects of the image.

At the very first stage of the processing hierarchy is the retina [19]. The retina senses visual information and transmits it to the brain via the optical nerve (approx. 1.5×10^6 fibers). While the number of fibers in the optical nerve is far beyond what we can replicate in an artificial system at the moment, it is far below the number of photoreceptors in the eye (approx. 10^8 receptors). If we further consider that some fibers respond only to motion and other transmit contrast rather than the photometric information of the receptors, it becomes obvious that this fascinating layer of neural tissue carries out some form of processing and data reduction. Indeed, the optical nerve fibers are axons derived from fourth or fifth order neurons in the

visual pathway [32], i.e. there are four or five layers of neurons processing receptors signals before the information is sent through the optical nerve.

The eye processes optical information even before the light is transduced into the neural signals; in addition to the lens focuses and the iris for rudimentary intensity adaptation, the photosensitive elements of the retina are spatially organized in a non–uniform way . The high spatial resolution of the fovea allows detailed sensing in the central region, while keeping a vague representation of the periphery of the image. The drawback of this strategy is the need for eye movement, which sequentially shifts the fovea to the "interesting" parts of the image.

In addition to these anatomical mechanisms for information compression, functional mechanisms exist in the higher processing centers of the brain. An example is attention — the ability to select a part of the retinal image to which the application of higher level processes can be restricted. Unlike eye movement, the attention shifts do not require any motor action, but occur internally, on a fixed retinal image. For this reason, attention shifts are faster than eye movements and appear to rapidly determine a number of interesting locations of the image. Then, the top–down pathways may initiate the eye movements for foveating onto one of these locations.

From this discussion it becomes evident that biological vision tightly couples sensing with processing and provides the top–down feedback for sensory adaptation and eye movement.

2 Computational Sensor Paradigm

Computational sensors [37] mimic biological systems: they incorporate computation at the level of sensing to improve performance and achieve new capabilities which were not otherwise possible. Computational sensors are usually VLSI circuits which may (1) include on–chip processing elements tightly coupled with on–chip sensors, (2) exploit unique optical design or geometrical arrangement of elements, or (3) use the physics of the underlying material for computation.

The computational sensor paradigm has potential to both reduce latency and facilitate top–down sensory adaptation, two main deficiencies of the computer vision at the moment. Namely, by integrating sensing and processing on a VLSI chip both transfer and computational bottlenecks can be alleviated: on–chip routing provides high capacity transfer, while an on-chip processor may implement massively–parallel fine–grain computa-

tion providing high processing capacity which readily scales up with the image size. In addition, the tight coupling between processor and sensor provides opportunity for a fast processor–sensor feedback for top–down adaptation.

3 Global vs. Local Operations

In the context of this thesis the global operations are important for two reasons. First, in perception it seems that each important decision is a kind of global, or overall, conclusion about a perceived world. These conclusions are often what a machine needs for coping with a task at hand. The global operations thus can be considered to produce the ultimate goals of the vision processing needed for the coherent interaction between a machine and the environment. Second, global operations produce *a few* quantities for the description of the environment. These quantities can be quickly transferred and/or processed to initiate an appropriate action for a machine. In addition, the results of the global operations can be used within the computational sensor in top–down sensory adaptation thus directing a further sensing and processing for more reliable performance.

Implementing global operations in parallel systems has been the subject of extensive research in both computer engineering and computer science. The main difficulty with implementing global operations comes from the necessity to bring together, or aggregate, all or most of the data in the input data set. This global exchange of data among a large number of processors/sites quickly saturates communication connections and adversely affects computing efficiency in parallel systems — parallel digital computers and computational sensors alike. It is not surprising that there are only a few computational sensors which implement global operations, all with modest capability and/or low resolution [18] [69] [70].

On the other hand, there are many computational sensory which implement local operations [7] [30] [38] [47] [49] [71] [77]. Those operations use only operands within a small spatial neighborhood of data and thus land themselves to the graceful implementation in VLSI. Local operations produce preprocessed images; therefore, a large quantity of data still must be read out and further inspected before a decision for an appropriate action is made — usually a time consuming process. Consequently, a great majority of computational sensors built thus far are limited in their ability to quickly respond to changes in the environment.

4 The Main Result

The primary aim of this thesis is to design computational sensors which reduce the latency in a vision system, and provide top–down feedback for more reliable performance. Such computational sensors must quickly provide reliable information necessary for coping with a task at hand. To attain this goal, this work embarks upon the problem or implementing global operations in computational sensors.

There are fundamental differences in how biological and artificial systems aggregate input signals. In digital systems, for example, gates with fan–in greater than 4 are rarely employed. The fan–in of an average neuron is 1,000 to 3,000, or even 10,000 [56]. Each input requires that a signal is routed to it. The more input signals, the more wiring is required, in both biological and artificial systems. Wires do not process information; therefore, economizing on wire should be important priority for both nerves and chips. Yet, the biological systems opted for large fan-ins. Some researchers [56] hypothesize that each neuron must have synaptic inputs representing *all* features that might ever be used, even though only a subset of them will contribute to any particular decision. Thus, it seems that the neurons are optimized for making global decisions about a large number of inputs, but using only a few of those inputs at a time.

In order to overcome obvious technological limitation for quickly communicating and processing large amounts of data, the proposed solutions draw upon the experiences of evolution and suggests the following implementation. The data are supplied optically by focusing an image (henceforth referred to as a retinal image) onto the array of photodetectors. A processor integrated within the chip, mimics neurons and makes a decision *based only on a few input data at a time*. The problem is how to efficiently chose which few input data to route to the global processor at each given time. This work proposes two mechanisms: (1) *sensory attention*, and (2) *intensity–to–time processing paradigm*.

4.1 Sensory Attention

The *sensory attention* follows the model of *visual attention* in brains. This analogy is attractive for 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 protect 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) [3]. Location of such attention must be maintained in the environmental coordinates; thus ensuring coherent behavior under ocular and head motion.

For implementation of attention several problems must be solved: (1) how to select interesting location within the retinal image, (2) how to shift the attention to another location, and (3) how to transfer data from focus of attention to the central processor for further inspection.

The winner–take–all (WTA) has been suggested for implementing location selection [43] [42]. The winner–take–all (WTA) mechanism determines the identity and magnitude of its strongest input [23]. We used a very compact VLSI realization of the WTA circuit originally proposed by Lazzaro [46] and Andreou [5]. The WTA uses a saliency map to guide the attention to the most conspicuous part of the retinal image. The saliency map can be derived from image features including the intensity, color, spatial and temporal derivatives, motion, and orientation. At the present state of technology we deliver the saliency map optically by focusing an image onto the array of photodetectors feeding the WTA network. This embodiment of the sensory attention we call *tracking computational sensor* because when the saliency map is a natural image, the trivial saliency map, the salient features are bright spots in the image and the sensor selects and tracks those locations.

We implemented the attention shifts by operating the tracking computational sensor in two modes: select mode and tracking mode (see Figure 1). In the select mode the sensor detects the global intensity peak within a programmable active region, a subregion of the retina. (This peak is called *a feature* in the context of the tracking sensor.)[1] The sensor continuously reports the position and intensity of the feature. By being able to program an arbitrary active region we ensure that the attention is directed towards parts of the image that are important for the task at hand. In the tracking mode the sensor dynamically defines its own active region, thus causing the sensor to ignore all retinal inputs except the currently tracked feature and its immediate neighborhood. This way our implementation ensures two things:

[1] In neurophysiology the pattern of activity which activates a visual neuron is called a *trigger feature*, a somewhat controversial notion. The area of visual field in which this pattern elicits the neural responses is called the *receptive field* of the neuron. Thus, in the context of the tracking sensor the sensor itself is a neuron whose trigger feature is the peak intensity within its receptive field (i.e. the active region.)

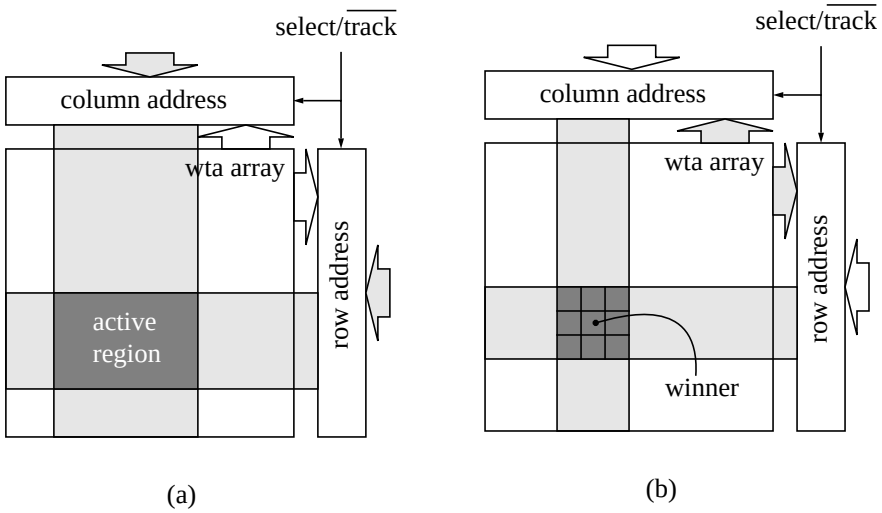

Figure 1: Modes of operation for the sensory attention computational sensor: (a) select mode, and (b) tracking mode.

(1) the location of attention is maintained in the environmental coordinates, (2) the sensor eliminates interference from parts of retinal image that are irrelevant for a particular task at hand. In the tracking mode, the sensor remains locked on the selected feature.

The WTA circuit reports the intensity of its winning input on a globally accessible wire. Therefore, by programming an active region consisting only of one cell (i.e. 1 by 1 active region), that cell becomes the winner and its intensity is reported. By scanning the trivial 1 by 1 active region around the attended location, the local data are transferred for higher processing.

The significance of our implementation of the sensory attention are summarized as follows:

- The global data — the position and intensity of the feature — are easily and quickly routed from the chip via several output pins.
- In the tracking mode these global data are also used internally for the self–defined active region. This represents an example of sensor/processor feedback presently missing in artificial vision systems. The tracking computational sensor demonstrates the significance of this feedback, as it is essential in preventing erroneous information from interfering with the currently attended salient feature relevant to a task at hand.

- In the select mode, the sensor can restrict its operation to an arbitrary size region of interest. In combination with a clever image formation, this renders the sensor useful in a range of practical applications.
- Inherent in our implementation is the ability of the sensor to provide random access to the image data if needed. The image data can be read from a random location within the retinal image including the vicinity of the feature being tracked.
- The size of a cell in a conventional 2μ CMOS technology is 62μ by 62μ, which is about equivalent to the area taken by a 4x4 pixel region in an industrial CCD camera. This is an appreciable spatial resolution, especially given the versatility of functions performed by the sensor.
- The dynamics of attention shifts have been found experimentally to range from 250 to 1,000 degrees/s. The attention shifts in humans occurs at a maximum of 125 degrees/s [57].

4.2 Intensity–to–Time Processing Paradigm

The other mechanism investigated is the newly proposed intensity–to–time processing paradigm — an efficient solution for massively–parallel global computation over large groups of fine–grained data [12]. Inspired by the human vision, the intensity–to–time processing paradigm is based on the notion that stronger inputs elicit responses before weaker ones. Assuming that the inputs have different intensities, the responses are ordered in time and a (global) processor makes decisions based only on a few inputs at a time. The more time allowed, the more responses are received, thus the global processor incrementally builds a global decision first based on several, and eventually on all the inputs. The key is that some preliminary decisions about the retinal image can be made as soon as the first responses are received. Thus, this paradigm has important place in low–latency vision processing.

The intensity–to–time processing paradigm was used to implement a *sorting computational sensor* — an analog VLSI sensor which is able to sort all pixels of an input image by their intensity, while the image is being sensed. In this realization the global processor essentially “counts” inputs (i.e. pixels) as they respond. The first input to respond receives the highest index, the next input one index lower, and so on. By the time all the inputs responded, the sensor has built an *image of indices*. The image of indices

represents the histogram equalized version of the retinal image. The two well know properties of such images are (1) the available dynamic range (of the readout circuitry) is equally and most optimally utilized, and (2) the image contrast is maximally enhanced. In many computer vision applications the histogram equalization is the first image preprocessing operation performed on camera images, primarily for signal normalization and contrast enhancement.

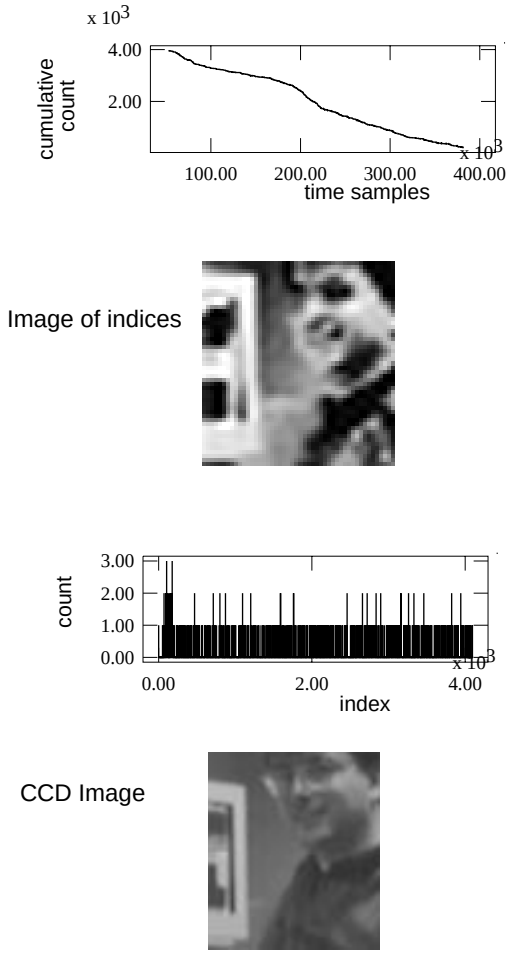

Figure 2: Scene 1 imaged by the sorting computational sensor. Top graph: cumulative histogram computed by the chip. Bottom graph: histogram of indices. CCD image is given to illustrate poor illumination conditions.

During the process of "counting" the global processor generates a waveform which is essentially the cumulative histogram of the retinal image. This waveform is one important global property of the retinal image which is

reported with low latency on one of the output pins before image is ever read out. Figure 2 shows one result of imaging with the sorting computational sensor. Figure 3 illustrates advantages of the imaging with sorting sensor over conventional linear cameras.

The significance of the sorting computational sensor are summarized as follows:

- The global information — a cumulative histogram of the sensed scene — is reported on an output pin with low–latency.
- This global information is used internally within the computational sensor to generate the image of indices. This is an example of the top–down processor–sensor feedback.
- The image of indices has uniform histogram; therefore, (1) the dynamic range of the output circuitry is most optimally utilized from information theoretic point of view, and (2) the contrast is maximally enhanced.
- Histogram equalization is often the first processing step in image processing. The sorting sensor preforms this operation in the analog domain at the sensory level. Therefore, the sensory signal suffers less noise corruption caused by the signal transfer and quantization.
- The image of indices never saturates. This is a better scheme for preventing saturation than the logarithmic photo detection proposed by other researchers [10] [56].
- The cell size of the sensor in 2μ CMOS technology is 76μ by 90μ. A sensor in 1.2μ CMOS technology is currently being fabricated with the cell size of 38μ by 38μ, which is about the size of a 3 by 3 pixel region in an industrial CCD camera. This is an appreciable spatial resolution for a sensor which implements a global operation on a massive amount of input data.

5 Future

It is generally believed that the future of computing depends on the exploitation of large–scale parallel processing. Although specialized parallel computers have been successfully used in many different application areas, there remain significant obstacles to the widespread use of parallel computers in task–oriented machine vision. The most significant obstacles

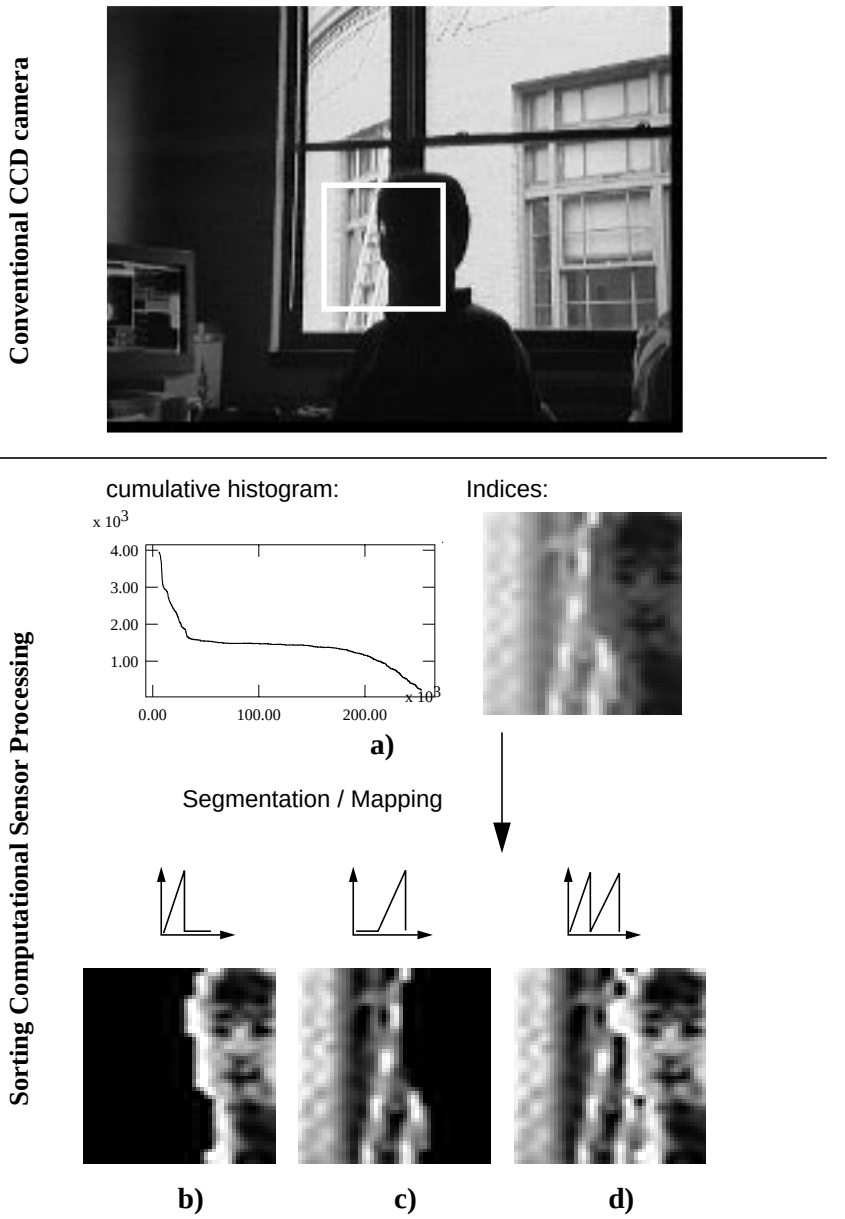

Figure 3: Sorting sensor processing: a) data from the sensors; b) segmentation (viewing the shadowed region); c) segmentation (viewing illuminated region); d) segmentation and shadow removal.

include the large size, power consumption and cost. The computational sensor proposed by this thesis are implemented in commodity VLSI technology. There is a strong indication that this technology will remain domi-

nant technology for many years. Furthermore, the cost of the technology will continue to go down, while its capabilities will continue to improve.

Tree–dimensional multi–chip packaging and through wafer interconnects are gaining increasingly more interest in VLSI community and will probably be widely available within few years. Then, one may imagine most of the low–level machine vision processing being implemented within a tree–dimensional stack of computational sensor chips. Many of these chips may implement various local operations as the information traverses through the stack of chips. Computational sensors performing global operations, however, will be essential at the higher levels of the stack. They will allow the results to be quickly routed off the stack for further high–level reasoning.

When this concept becomes possible, the low–latency, robust performance, low power and portability will make many new applications for machine vision possible. The area which will probably be the most dramatically impacted is a human–machine interaction. Humans will start seeing increasingly more vision based systems *around* and *on* themselves: in their homes, cars, offices, hospitals, entertainment, computers, etc. Therefore, the future of computational sensor seems promising. The low–latency computational sensors performing global operations on massive amount of data will find important place in that future.