Pergamon

Neural Networks, Vol. 8, No. 7/8, pp. 1297–1308, 1995
Elsevier Science Ltd
Printed in Great Britain
0893–6080/95 $9.50+.00

0893-6080(95)00092-5

1995 SPECIAL ISSUE

Space-variant Active Vision: Definition, Overview and Examples

Eric L. Schwartz, Douglas N. Greve and Giorgio Bonmassar

Boston University and Vision Applications Inc.

(*Received and accepted* 9 *May* 1995)

Abstract—The term space-variant vision was introduced in the late 1980s to refer to sensor architectures based on a smooth variation of resolution across the workspace, like that of the human visual system. The use of such sensor architectures is rapidly becoming an important factor in machine vision in which the constraints of size, weight, cost and performance must be jointly optimized. The structure of this paper consists of four parts. A review of the four generic architectures for vision will be presented, providing a context for the term "active vision", and a justification for the importance, and the connection between, space-variant architectures and active vision methods. A brief quantitative review of the specific space-variant properties of primate visual cortex topography will be provided, in the context of sensor design. The engineering and algorithmic problems that are associated with exploiting space-variant systems will be stated. Examples of several recently constructed miniature space-variant active vision systems will be briefly reviewed, along with a brief discussion of solutions to the basic problem areas in space-variant vision.

Keywords—Fovea, Space-variant, Active vision, Pyramid, Computer vision, Visual cortex.

1. INTRODUCTION

The term space-variant vision was introduced in the late 1980s (Schwartz et al., 1988; Yeshurun & Schwartz, 1989) to refer to sensor architectures based on a smooth variation of resolution across the workspace, like that of the human visual system. An alternative term that is also in use is "foveating vision systems". Space-variant architecture shares many features, and some important differences, with the "pyramid architecture" that is more familiar in computer vision.

The use of such sensor architectures is rapidly becoming an important factor in machine vision. Burt's truncated pyramid (Burt, 1988) and other, more specifically space-variant applications (Sandini & Dario, 1989; Sandini et al., 1989; van der Spiegel et al., 1989; Weiman, 1988, 1990; Baloch & Waxman, 1991; Bederson et al., 1992; Engel et al., 1994) have been described. Based on theoretical analyses, which will be summarized below, and from our experience over the past 6 years in building several state-of-the-art space-variant machine vision systems, we believe that space-variant vision is the key to producing small, high performance, light-weight and inexpensive computer vision systems.

The main areas of application in which the constraints of size, weight, cost and performance must be jointly optimized include:

- Military applications in which target identification and classification from a wide field of view must be performed by a small, low power system and communicated to a human user.
- Applications in which a visible or IR camera system is to be used to analyze a large work-area, and communicate the scene interpretation to a human observer via non-visual cues.
- Surveillance applications for public spaces (e.g., intelligent highway applications, factories, military installations), and private spaces (e.g., monitoring vehicles, homes, etc.),
- Autonomous and teleoperated vehicle control.
- Image communication over limited bandwidth channels, such as voice-band telephony.
- "Wearable" prosthetic camera-systems for the blind and visually impaired.

Acknowledgement: Supported by ARPA ANNT Program, ONR Contract # N00014-92-C-0119.

Requests for reprints should be sent to E. L. Schwartz, Boston University, Department of Cognitive and Neural Systems, 677 Beacon Street, Boston, MA 02215, USA.

The broad range of applications claimed here is indicative of the large segment of machine vision and visually guided robotic applications that are constrained by joint size, cost, weight, and performance issues. Although this is an obvious statement, the less obvious conclusion, which will be supported in this paper, is that space-variant active vision, based on the architectural principles of the human vision system, provides the best route to this joint optimization. This is because a "lever-arm" exists which is, for biological systems, as large as four orders of magnitude (Rojer & Schwartz, 1990), and for the systems which will be reviewed in this paper (Bederson et al., 1992; Engel et al., 1994) is as large as two orders of magnitude. This claim, which is based on two decades of research (for a recent review, see Schwartz, 1994) into the quantitative nature of space-variant vision in humans, and the construction of two generations of space-variant computer vision systems, will be established in this paper. Its significance for the neural net ATR community is twofold:

1. The connection between the architecture of primary visual cortex and practical "design wins" in computer vision provides one of the most prominent links between neuroscience and realized (as opposed to promised) engineering applications.
2. The 2–4 orders of magnitude advantage provided by these architectures translates directly into similar advantages in size, cost, and weight advantage, and suggests the possibility of "commodity robotics", a term introduced by Bederson et al. (1992) which refers to a radical drop in the cost of visually guided robotic applications. Just as the introduction of "commodity computing", in the form of the IBM PC in the early 1980s led to a major societal transition in the availability, use, and application of digital computers, the term "commodity robotics" suggests the possibility of a similar transition towards the widespread use of machine vision technologies which have, until now, been restricted to high-end application domains. This transition has not yet happened, and requires a radical reduction in the size and cost of machine vision technologies. It is the suggestion of this paper that space-variant active vision will supply the basis for this transition.

The structure of this paper consists of four parts.

- First: a review of the four generic architectures for vision will be presented, providing a context for the term "active vision", and a justification for the importance, and the connection between, space-variant architectures and active vision methods.
- Second: A brief quantitative review of the specific space-variant properties of primate visual cortex topography will be provided, in the context of sensor design.
- Third: The engineering and algorithmic problems that are associated with exploiting space-variant systems will be stated.
- Finally, examples of several miniature space-variant active vision systems will be briefly reviewed, along with a brief discussion of solutions to the basic problem areas in space-variant vision.

2. FOUR ARCHITECTURES FOR VISION

In the following discussion, the term "active vision" will be defined in terms of a biological or machine vision system in which the sensor (i.e., retina or solid-state imaging chip) is moved via robotic or muscular actuators. Human vision is active. Computer vision systems are active if, and only if, they are actuated.[1]

The terms "active vision" and "foveating" vision systems are sometimes used to describe passive systems, i.e., systems which are not actuated, but in which a simulated fixation point is manipulated via software. It should be evident that a passive system should not be called "active", but precision of terminology is often violated in this area. The following nomenclature is offered as a means of classifying possible visual architectures, in the hope that a uniform terminology can be used to accurately classify the possible range of visual architectures.

1. *Space-Invariant Passive Vision: SIPV.* The SIPV architecture is currently the dominant one in machine vision, accounting for all but a tiny fraction of currently deployed systems. In this architecture, the visual field is sampled uniformly, i.e., by constant sized pixels, and the camera is stationary. The advantage of this architecture is that few engineering problems are

[1] Given the fact that human vision is active, and that human vision provides the implicit definition of the term vision, as opposed to a more neutral term such as multiple dimensional signal processing, it would not be unreasonable to use the term vision to be mean, implicitly, "active-vision", while reserving the term "passive vision" for the currently dominant computer vision architecture of passive vision(!) In fact, by this same train of thought, it would not be unreasonable to define the term "vision" in terms of the "space-variant active" architecture that is universal in the higher vertebrate biological systems, and use the nomenclature which is presented in this section to characterize the other three possible architectures.

involved in its application. "Off-the-shelf" cameras, no robotic actuation issues, and the use of standard algorithms make the SIPV architecture the one of choice for most computer scientists. The disadvantages of this architecture is that the space-complexity[2] is unfavorable. The number of pixels increases quadratically with the angular size of the work space and resolution of the sensor. The unfavorable space-complexity of the SIPV architecture is one of the principal blocks to progress in building small, inexpensive, high performance vision systems. A biological example of an SIPV system is approximated by that of the goldfish, in which visual resolution is roughly constant from center to periphery, and in which eye-movements are relatively unimportant.[3]

2. *Space-Invariant Active Vision: SIAV.* Adding a pan-tilt motor to a conventional TV sensor defines the SIAV architecture. The advantage of this architecture is that the work-space is increased by the angular swing of the actuators, at no extra cost in space-complexity, i.e., for the same burden of pixels/frame. The majority of "active-vision" applications are SIAV (e.g., Clark and Ferrier, 1988; Ballard, 1990; Fiala et al., 1994). For a 50° camera lens swinging through a 150° work space, the advantage in space-complexity of the SIAV approach is about an order of magnitude (i.e. nine time the work space for the same number of pixels). The disadvantage of this approach is the need to deal with expensive, bulky, and (for many computer scientists) unfamiliar issues of robotic actuation. In fact, it would appear that the introduction of mechanical elements (e.g., a pan-tilt actuator) to an otherwise purely electronic system would be a weak-point of this approach. Many have argued that increasing the size and read-out speed of conventional TV sensors is a more sensible approach, using electronic panning and scrolling to avoid the introduction of mechanical actuators. On the other hand, it should be noted that the hard-disk drive, which is based on robotic and mechanical principles, has to his day remained cost-effective with respect to purely electronic storage (e.g., flash memory, DRAM, etc.). Nevertheless, the SIAV architecture is relatively uninteresting, providing no more than an order of magnitude advantage in space-complexity, and hence vulnerable to advances in sensor technology and evolution to higher density sensor arrays.[4]
3. *Space-Variant Passive Vision: SVPV.* Forveating systems without actuators have been used, for example by Burt (1988). This type of system, however, is severely constrained by the size of available sensors. It is necessary to have a very large sensor (e.g., 2000×2000) to be able to electronically pan and tilt a 512 × 512 frame over a reasonable workspace. The cost of megapixel sensors is currently prohibitive, and will always be relatively high due to the general rule that VLSI scales in cost exponentially with its area. The advantage of this architecture is that it is entirely electronic: no robotics are involved. The disadvantage is that it appears to be intrinsically cost-ineffective to use extremely large sensors, leading both to high sensor cost, and to high cost for the CPU and memory requirements associated with the manipulation of the megapixel arrays that would be required to allow a static camera to electronically pan and scroll through a large workspace at high resolution.
4. *Space-Variant Active Vision: SVAV.* All higher vertebrate vision systems (cat, owl, hawk, monkey, human) are SVAV systems, utilizing one (cat, monkey, owl, human) or more (e.g., hawk) "foveal" areas to achieve a large workspace and high resolution without incurring the burden of a huge number of "pixels". Non-uniform sampling of the image is the key defining idea, and a non-uniformly sampled image implies the need for actuators, or some other means (see SVPV systems above) to "point" the sensor. This in turn implies the need for sophisticated "attentional" algorithms to guide the actuators of the system, and also implies the need for image processing algorithms which will likely be quite different from standard space-invariant approaches. The disadvantages here are obvious: there are difficult engineering and algorithmic problems associated with the SVAV approach. For this reason, no more than a handful of labs have produced working systems based on SVAV architectures (e.g., Weiman, 1988, 1990; Sandini & Dario, 1989; Sandini et al., 1989; van der Spiegel et al., 1989; Baloch & Waxman, 1991; Bederson et al., 1992;

[2] See Rojer and Schwartz (1990) and section below for a definition of the term space-complexity.

[3] However, in the lower vertebrate and invertebrate systems, which lack high evolved opto-motor systems, head and body movement almost certainly supply some aspect of active vision.

[4] One advantage that has been suggested for active vision in general, which would apply to SIAV systems, is that ill-posed problems in vision might be solved by use of the multiple views of a scene provided by an active vision system (Aloimonos et al., 1988). It is difficult to quantify this idea, which undoubtedly has merit. On the other hand, it would seem to be a commonplace introspective observation that, provided we are looking in the right direction, a single glance provides a good solution to many vision problems, suggesting that the "ill-posed problem" advantage of active vision is incremental, rather than fundamental.

Engel et al., 1994). The advantages of this approach are that up to four orders of space-complexity advantage may be achieved with SVAV architectures (see next section for detailed analysis). This is beyond doubt the reason that all higher biological vision systems use this approach. The principal engineering difficulties in building miniaturized SVAV systems have been solved, as illustrated by the proof of concept CORTEX-I and CORTEX-II systems, to be described below.

Of the four basic architectures for vision, current engineering approaches are strongly clustered in the category of SIPV, while biological vision, at the high end of evolution, is entirely represented by the space-variant active vision (SVAV) architecture. The space-variant approaches to machine vision that have been developed tend to be in the category of "pyramid" architectures, which, in the present context, is more accurately described as "multi-resolution" than fully space-variant. Only a small number of research groups have investigated fully space-variant active vision applications in machine vision at the hardware level.

At the present time, we can only guess what the future of machine vision will be, vis-à-vis sensor architecture. However, a brief discussion of the issue of space complexity will tend to suggest that SVA architectures are likely to dominate the future of light weight, low cost systems, based on the following analysis.

3. SPACE COMPLEXITY IN MACHINE VISION AND TEN THOUSAND POUNDS OF BRAIN

The human visual system is able to cover a wide visual field, and achieve high maximum resolution, without the need for an unreasonably large number of spatial channels. The foveating, or space-variant architecture of visual cortex provides a dramatic form of data compression. Just how dramatic this compression is can be seen from the following simple estimate. Suppose we wish to "cover" a solid angle which is roughly comparable to human vision (let us say about 100°×100°),[5] with the same maximum resolution as that of human vision, which is about 1 minute of arc. If we were to attempt this with conventional video sensor technology over a 100° × 100° field, we would require our sensor to have $6000^2 \times 2 \times 2$ pixels (the factor of 2 × 2 is for sampling, based on Shannon's theorem. Actually, practical oversampling rates are considerably lager than this minimum estimate, and the extent of the human field is also considerably larger than 100° × 100°). The simple "back-of-the-envelope" estimate outlined here leads to a pixel count of 1.44×10^8 per frame.

In a more careful analysis of this compression factor, Rojer and Schwartz, (1990) have defined a measure of sensor quality, which was termed F/R quality, defined as the ratio of sensor field of view to maximum resolution, as outlined above. This is a measure of the spatial dynamic range, or space-complexity, of a sensor.

In Rojer and Schwartz (1990), it is shown that for a given F/R ratio, which is taken to be an estimate of the quality of the vision system, the asymptotic space-complexity of the SIPV (and SIAV) architectures scale quadratically with the rank of the sensor matrix. In other words, to double the F/R ratio for these architectures requires four times as many pixels. The SVAV architecture, however, has logarithmic asymptotic space-complexity. Thus, to double the F/R of an SVAV sensor requires an increase of log 2. This is an outstanding property of the SVAV architecture, which is reminiscent of the computational advantage of the FFT versus the DFT, which have asymptotic computational complexity for a two-dimensional (i.e., image processing) FT problem that is $O(N^2 \log N)$ and $O(N^4)$, respectively, where N is the rank of the sensor array.

The space-complexity of a vision system is a good measure of its computational complexity, since the number of pixels which must be processed is determined by the space-complexity. Thus, even though the space-complexity does not entirely determine the computational complexity (which depends on the specification of one or more algorithms), we believe that the space-complexity is a good measure of the computational complexity, and will, in fact, likely be proportional to it. To make this analysis more concrete, consider the following numerical estimate: considering the primate visual field to be 140° (vertical) and 200° (horizontal) and using current estimates of the topography of human visual cortex, the number of "pixels" in a complex log sensor such as the human visual system is estimated by Rojer and Schwartz (1990) to be about 150,000. This number is consistent with the number of fibers in the optic tract (about 1,000,000), since we have not accounted for color, on-off and off-on pathways, non-cortical afferents in the optic tract, and redundancy of sampling. We believe that a count of about 10^5 "pixels", or "sampling units" or "spatial degrees of freedom" is consistent both with cortical topography and the number of fibers in the optic tract[6].

[5] The human visual field actually covers a range that is roughly 180–220° horizontally by 140° vertically, for both eyes.

[6] Nakayama (1990) has also provided a estimate of the number of "pixels" required to encode contrast. He obtained an estimate of 25,000 "pixels" somewhat lower than ours, but he did not provide full details of his calculations such as assumptions on sampling, parameters of human topography, etc.

The number of pixels in a conventional, space-invariant sensor (e.g., a TV sensor) of the same F/R ratio, is 600,000,000.[7] These estimates for the space-variant and space-invariant pixel burden of vision sensors suggest that compression ratios of between 3500:1 and 10,000:1 are achieved.

Since the primate cortex is roughly 50% (exclusively) visual, and the human brain weighs about 3 lbs, it seems clear that our brains would weigh many thousands of pounds if we were to maintain the same spatial dynamic range, but used a space-invariant, or non-foveal architecture.

Since wide angle vision with high acuity would appear to be of great selective advantage, and since a brain which weighs 5000–30,000 lbs is not, it appears that we have identified at least one indisputable functional correlate of visual cortex spatial architecture.

4. ENGINEERING AND ALGORITHMIC PROBLEM AREAS IN SVAV

The previous two sections of this paper have provided a nomenclature for vision architectures, and a brief outline of the space-complexity issues which favor the SVAV architecture for applications in which size, cost, weight, and performance must be jointly optimized. However, the use of the SVAV architecture in ATR and other application introduces a number of difficult engineering and algorithmic issues. Some of these issues are generic to all real time applications, e.g., the choice of computer platform and development environment. But others are specific to the SVAV architecture, requiring novel solutions to difficult problems in computer science and engineering. In this section, a brief review of the general problem areas of SVAV will be provided, and in the final section of the paper, recent specific solutions to these problems will be outlined via a demonstration of two recent SVAV systems: CORTEX-I and CORTEX-II, which have been built by Vision Applications, under contract to ARPA.

The first problem area in SVAV is access to space-variant image formats at video data rates. This problem is compounded by the fact there are no space-variant sensors commercially available (at the time of writing this paper), although one academically oriented group has produced, and offers, a workable space-variant CCD imaging chip (van der Spiegel et al., 1989).

There are two generic approaches to providing space-variant sensing. The first, as just mentioned, is to fabricate custom VLSI sensors which are intrinsically space-variant. The second route is to use a computer synthesis of a space-variant scene, via a conventional "off-the-shelf" sensor chip, using special or general purpose hardware to "warp" the output of a conventional sensor chip to the form of a space-variant output, e.g., that of the log-polar mapping. Baloch and Waxman (1991), and Weiman (1989) have used the PIPE, which is a high-speed pipe-lined image processing accelerator, to form space-variant images at video rate. One drawback of this approach is that special purpose devices such as the PIPE are large, expensive, and somewhat inflexible in their programming capabilities.

Our group has followed both approaches (custom VLSI and dedicated hardware acceleration of a conventional sensor), but has much greater success with the latter. The reason is that sensor evolution is very rapid, and very expensive. University or academic design and fabrication of sensors tends to result in sensors which do not have the visual performance of good commercial sensors, and academic research necessarily lags behind industrial research, for simple reasons of access to infrastructure and financing. We have developed highly optimized algorithms for synthesizing space-variant frames from conventional commercial sensors, and we thus find that the advantages of using commercial high quality sensors, with a fast algorithm that is implemented via a DSP or gate array, is the most favorable route to space-variant vision. In the SVAV systems CORTEX-I and CORTEX-II, we have been able to synthesis space-variant (log-polar) images, using a single DSP (either AD2101 or TI320C40) and minimal memory (less than 10K bytes) at 30–50 frames/s. This requires highly optimized algorithms and code, but in the end provides the most efficient route to providing space-variant image data at frame-rate. Also, aside from allowing us to exploit the latest and best commercial sensor technologies, this route also has the advantage of allowing a flexible choice of map parameters. Hardware sensors have fixed geometries, while synthesized space-variant sensors may be changed for different problems, and, in fact, may be changed in real-time, at frame rate.

4.1. Actuation: Spherical Pointing Motor

Actuation, in so far as it is intrinsic to the definition of active vision, is clearly a problem, since high-

[7] Shostak (1992) has estimated the following "pixel" estimates for the full visul field, using a space-invariant (non-foveal) architecture:

- Solid angle of human vision: 15,000 degrees^2 (180 deg(horizontal)×135 deg(vertical).
- Max. resolution: 0.5 minutes of arc.
- Sampling factor: 2.
- Space-invariant sensor size: 36,000×28,000= 1,000,000,000 pixels.

Shostak's estimate is larger than ours because he used an assumed 0.5 minute of arc, rather than a 1 minute of arc maximum resolution.

speed, high-accuracy actuation is required, ideally in a form which is lightweight and inexpensive. It is common to see active vision systems which are mechanically compromised, such that the response time of the entire system is markedly slowed down by mechanical resonance in the actuator system. Equally common are systems which are "hacked" together from low-cost airplane servo-motors. Although small and inexpensive, the end-result of this strategy usually ends up being surprisingly large, and the inherent inaccuracy of hobby style servos tends to provide a low-performance system. At the opposite extreme are systems built from high-quality DC servo-motors (e.g., Clark & Ferrier, 1988; Fiala et al., 1994), with care paid to mechanical engineering and control, but which tend to be large, heavy and quite expensive.[8]

The reason for this is that current actuators are generally based on DC motors which have a single degree of freedom. Multiple degree of freedom systems thus must be built up from multiple motors. This, in turn, leaves two generic choices.

Direct drive. A direct drive system may be built, in which one motor "rides" on another. This "motor-on-motor" design is problematic, since the inertial constraints that are inherent in high-speed precise actuation force one of the motors to be quite large [see Bedersen et al. (1994a) and Greve (1995) for discussion]. Also, high performance direct drive DC servo motor systems tend to be expensive and bulky.

Linkage actuator. A linkage system may be built with two independent motors of comparable torque. Systems following this design generally are larger and less accurate than direct drive motors.

An alternative solution is to build a direct drive DC actuator which has multiple degrees of freedom inherent in its design. In the next section, we will describe such a novel actuator, called the spherical pointing motor (SPM), which has been developed by Vision Applications, Inc. for the specific purpose of actuating miniature high performance SVAV systems. The SPM is the only actuator to date which jointly satisfies the criteria of high-speed precision, small size, and low cost. The next section of this paper describes the SPM in more detail.

In addition to the problem areas already discussed, there are a number of technical problems in building small, high performance vision systems which are generic to machine vision in general. These include issues of computer platform, power systems, development environment, real-time programming, and image processing and pattern recognition. Two special areas need to be emphasized, which are "attention", since space-variant systems cannot function without an effective solution to the question of "where-to-look-next", and image processing, which tends to specialized, and much more difficult, when performed on a space-variant architecture. The final section of this paper will review all of these problem areas by means of presenting solution examples from the systems CORTEX-I and CORTEX-II.

5. TWO SVAV SYSTEMS: CORTEX-I AND CORTEX-II

The space-variant or foveating architecture of the primate visual system has recently begun to be applied to the construction of high performance machine vision systems.

During the past several years, with support from ARPA's Artificial Neural Network Technology program, Vision Applications, Inc. has built two machine vision systems (CORTEX-I and CORTEX-II) which utilize the complex log geometry as its sensing strategy. The systems that have been constructed have established extremely high performance on certain measures, which will now be reviewed.

CORTEX-I was completed 2 years ago: it is a minature space-variant active vision system based on a complex logarithmic sensing strategy (Bederson et al., 1992; Wallace et al., 1994). The benchmark application for this system was to acquire moving targets (automobiles), track them with the camera, and to use pattern recognition techniques to read the license plates of the cars as they drove past the camera system. The choice of the license plate problem was that it provided a generic set of problems (sensor design, actuation, attention, classification, image processing), which were used to develop the basic principles of this type of system, but which could easily be generalized to other domains.

In order to solve this problem, a series of hardware and algorithmic problems were solved. At the hardware level, a novel actuator design was produced and implemented, called the spherical pointing motor. This design produced a two-degree of freedom camera pointing device which was fast,

[8] We constructed a system from commercial DC servi-motors obtained from Klinger, Inc. The system performance was excellent, but it was extremely expensive to build (roughly $15,000 for two axes of control, for actuators and electronics), and quite large and heavy. An added drawback of this system, as is the case for other high-performance systems based on conventional actuator technology, is that they are physically dangerous since they possess sufficient torque to break the arm of a careless experimenter.

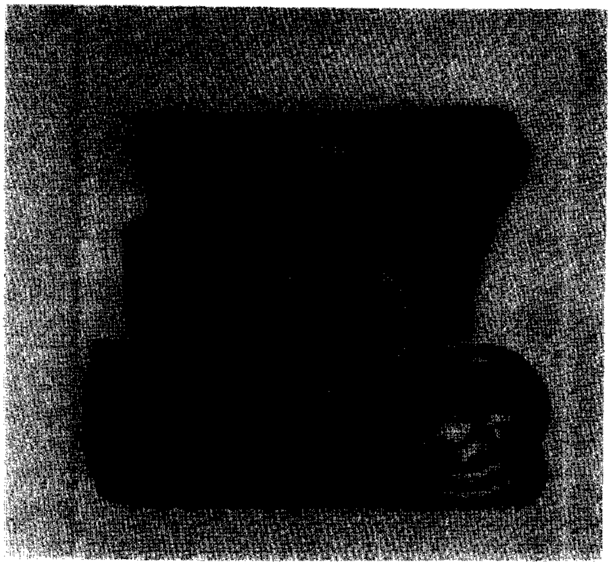

FIGURE 1. First prototype spherical pointing motor.

FIGURE 2. Cortex-I active vision system, including camera, motor and computer control and image processing.

very compact, and extremely inexpensive to produce. An early SPM prototype is shown in Figure 1.

Additional hardware innovations were the production of sub-miniature camera and lens systems, and custom VLSI Sensor and DSP based image processing hardware. An example of a fixed focus miniature camera is also shown in Figure 1.

At the algorithmic level, it was necessary to develop attentional algorithms that were capable of locating an object of interest (e.g., a license plate) in a complex scene, containing moving objects, in real time, to track the object of interest, and to perform image processing and pattern recognition on the tracked object. This work is fully described in a recent series of papers (Bederson et al., 1992; Ong et al., 1992; Engel et al., 1994; Wallace et al., 1994).

Figure 2 shows the hardware platform of DSP and micro-controller chips that made up CORTEX-I's platform. The license plate reading benchmark was achieved (Bederson et al., 1992) with a hardware system that occupied less than 0.5 ft^3, weighed less than 10 lbs, and cost roughly $2000 in parts to build, inclusive of video camera, lenses, motors, and computer system.

A sample of the space-variant images in the license-plate reading task is shown in Figure 3. The most notable aspect of this system was its ability to perform a difficult machine vision task, in real time, with the support of only 12 MIPS of processing power. This system is roughly 10–100 times smaller, cheaper, and computationally less expensive (in MIPS) than other contemporary machine vision systems. The reason for this economy can be traced directly to the use of a space-variant sensing strategy. The system processed only 1400 pixels per frame, instead of the usual 64,000–256,000 pixels common in machine vision. In effect, we exploited the same type of leverage, outlined above for the human visual system (although not quite the same magnitude) and the scaling down of our systems cost and size

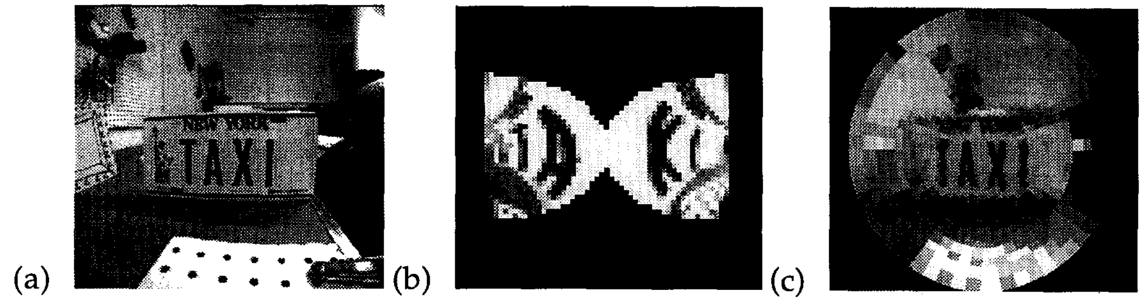

FIGURE 3. (a) License plate image at 256×256 resolution; (b) Logmap, simulating planned Synaptics Sensor (1200 pixels); (c) Inverse logmap, or retinal view of Figure (b).

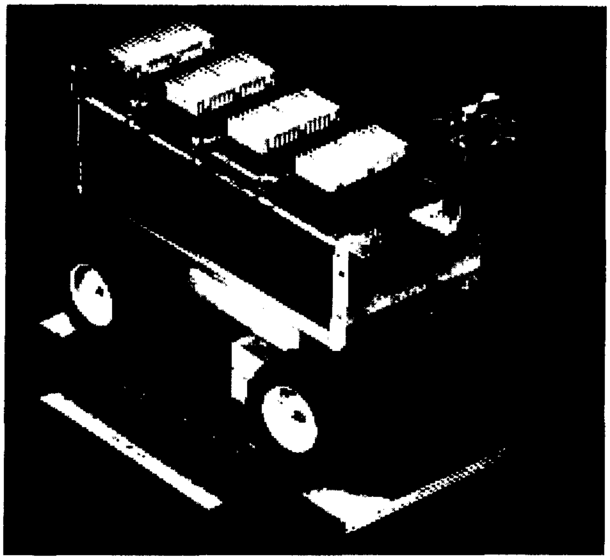

FIGURE 4. Cortex-II. A 12 in rule is provided for scale in the lower left. The SPM camera system is mounted on the upper right corner of the vehicle. Wire wrap boards containing vehicle and actuator control latches, frame grabber and motor control are seen as vertical cards within the VME 3U card cage. The C40 DSP's and an Air-LAN wireless ethernet board are mounted vertically behind the card cage, and an embedded PC and disk drive are mounted under the card cage. A nickel metal hydride (NIMH) battery is seen under the card cage on the left side.

must be understood to apply not only to the sensor, but to the memory, the CPU power, and to almost all other aspects of the system.[9]

A second generation system, CORTEX-II, is currently being completed, under support from ARPA, which will provide 200 MIPS or more, and which is mounted on and controls a miniature robot vehicle. The benchmarks of this project are autonomous driving, target acquisition and classification, and teleoperation via wireless links to the vehicle over the existing cellular telephone network. Figure 4 shows the status of the CORTEX-II vehicle at the present time.

5.1. Design and Fabrication of Camera and Optics

The camera system shown in Figure 5 implements a miniature, actuated zoom lens. In previous work, we found that the presence of a zoom lens was a significant help in performing pattern recognition tasks, and that the advantages of the space-variant architecture work very well with a zoom capability. In

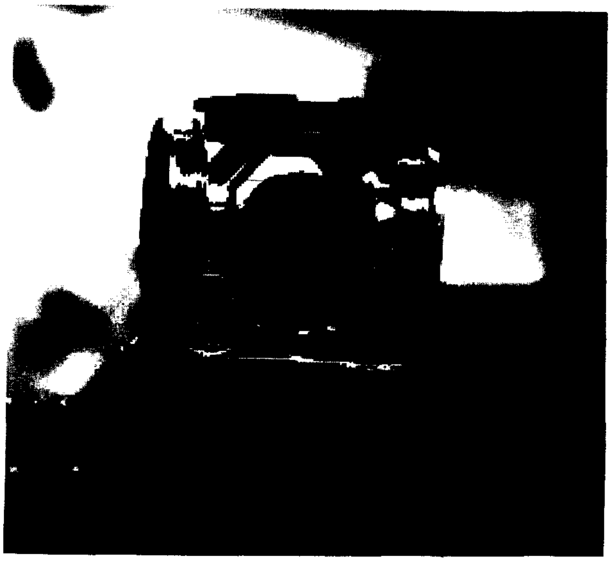

FIGURE 5. Spherical pointing motor active vision camera. A US Quarter is shown for scale. Four of the five drive coils are visible here, and the lens and magnetic rotor are visible in the center. The bearing is a simple pin bearing gimbal design. The lens is an actuated zoom lens with a small coil on the lens acting as a solenoid in the magnetic field of the main magnet rotor.

our current system, we have implemented a miniature zoom lens which is actuated by a small coil attached to a single transfer lens. When a current is applied to this coil, it is actuated by the stray magnetic field of the main rotor magnet. Thus, we have been able to add zoom lens functionality with a minimal increase in the size and weight of the system.[10]

5.2. Actuation: Spherical Pointing Motor

Actuation of the camera is a critical detail of any active vision system. If the system is to be small, lightweight, and high performance, the availability of "off-the-shelf" actuators is problematic. We have developed a novel actuator, called a spherical pointing motor, originally described in Bederson et al. (1992, 1994b). This device uses a system of three orthogonal coils to point a rare-earth rotor, upon which the sensor and optics are mounted.

Solutions to the problem of the control of this type of actuator are described in Greve et al. (1994). The dynamics of the SPM are comparable to those of the

[9] Like the human eye, the log map sensor does not require high quality optics off-axis, as conventional cameras do. This made possible very small and light lenses, which in turn allowed actuators to be very small and light. There are a number of synergistic benefits which followed from the complex log sensor geometry.

[10] Naturally, biological systems do not use zoom lenses. However, a system such as that of the human has an effective resolution that would be equivalent to a 16,000×16,000 sensor (see Rojer and Schwartz, 1990). The extreme spatial dynamic range of the human sensor thus makes a zoom lens somewhat redundant. However, for machine vision applications, where the initial sensor is equivalent to a range that is more like 512×512, zoom capability becomes a much appreciated "boost" that is critical for fine grained pattern recognition applications.

human eye (saccadic speed of about 600°/s), while the size and weight of the current SPM camera system, which measures 5.8×5.8×5.8 cm and weighs 140 g, is close to that of the human eye, especially in comparison with other contemporary active vision actuation systems.[11]

5.3. Attention

The use of SVAV methods requires an attentional algorithm, since the sensor is useless unless the foveal region can be "pointed" at a region of interest, in the absence of a final analysis of the scene, which will only be complete once the scene is fully analyzed. This obviously introduces an aspect of "recursion" to the use of SVAV systems, with the interaction of the attentional algorithm, and the final stages of pattern classification, involved in a form of recursive loop.

The definition of attention is itself a problematic issue, since the term is used in a wide variety of contexts. There is an enormous amount of work in the psychological literature that would appear after a literature search under "attention". For our purposes, we find that virtually none of this psychological (or physiological) literature is of much direct relevance to the technology of space-variant systems. The reason for this is that there is really only one difficult problem to be solved, and this problem has not been directly addressed in the biological or psychological literature: it is the problem of "resource allocation". In other words, we must find a way for a sensor system to determine, in a model based context, where a target of interest is likely be located. This determination is intrinsically unreliable, since, by assumption, the target is likely to lie in a region of the sensor field which has low resolution. Thus, we need to find a fast and cheap method which will locate likely regions of interest, use this inference to point the sensor, and then continue the process. A successful algorithm is one that "usually" completes this process in no more than a few fixations. We cannot expect a "single-fixation" solution in general, and we cannot tolerate a solution that takes more than a few fixations.

When expressed in this way [see Yeshurun and Schwartz (1989) for an early discussion of attention in the context of SVAV], it becomes clear that "attention" is merely a form of "fast and dirty" pattern recognition. Moreover, an SVAV system that satisfied this definition would perform reasonably well in most practical applications.

We have developed an approach to attention that builds on this idea [see Rojer and Schwartz (1992) for a detailed presentation of this algorithm]. Briefly, we define the attention problem to be a projection of high-dimensional feature and object spaces onto a low dimensional "fixation" space, i.e., a space of two dimensions: azimuth and elevation of the sensor optical axis.

Finding the high-dimensional correspondence between features and objects is the usual pattern recognition problem, and is solved by the end of the series of fixations. However, initially, we only wish to find equivalence classes of "interesting" directions to point the sensor. This is done by pre-computing feature–object relationships off-line, and then using the feature cues alone, at run-time, to project directly to the two-dimensional space of fixations. This is fast: we do not need to consider object structure at all at run time. It is "dirty", since we obtain "equivalence classes" of objects, represented by their location in space, rather than object classifications themselves. In practice, we have made this algorithm the basis of our SVAV attentional algorithm, and it has worked quite well on the model-based problems that we have investigated (e.g., the license plate problem). We expect to continue to use it in the future, and feel that it provides a general basis for providing attention, in a model-based context, for SVAV systems.

5.4. Computer Platform

Four parallel, floating point DSPs (Texas Instruments 320C40) provide the computational engine of CORTEX-II. The C40 chip has six high-speed parallel ports (called comm-ports) which are capable of a 20 megabyte/s input-output. The comm-ports allow the parallel DSPs to connect to each other, and to external peripherals such as the camera, actuators, etc., via high-speed I/O links. Although many have found the comm-ports difficult to work with, due to the high-speed "handshaking" required on transmission and reception of data, we have found that Alterra EPLD devices can be used to interface slow external devices to the high-speed comm-ports with good results.

The parallelism supported by the C40, together with its high-speed DSP instruction set, makes it an excellent, and at the present time, unrivalled, choice for the basis of a small high performance vision system.[12]

[11] An additional advantage of the SPM is that it is constructed from a few hundred feet of magnet wire, a small rare earth magnet, and a simple pin-bearing gimbal. No precision machining, encoders, or other expensive components are required, and we estimate that the actuator alone could be built for less than $10.

[12] Recently TI has announced the C80 processor, and Analog Devices has announced the 21060 SHARK processor. Both of these devices are extremely interesting, but we view them as immature at the present time, particularly vis-à-vis software support, for use in the next year or two.

5.5. Power System

Power is supplied to the system by a set of Energizer Hydritech Nickel–Metal Hydride (NiMH) batteries coupled to Vicor DC-DC converters. The Vicor modules can maintain their output voltage for input voltages varying from +10 to +20 V. We use a watchdog/cutoff circuit to shut down the system if a critical condition is detected. The watchdog circuit monitors all battery voltages and a request from the user to halt the system. The power supply is designed to operate with batteries or an external power source and can be used to charge our batteries without disconnecting them from the robot.

CORTEX-II requires 60 W on the +5 V supply, 48 W of +12 V power, and 40 W of −12 V power. Using NiMH batteries and DC-DC converters we can operate our "development" prototype for up to 1 h.

5.6. Development Environment

The software development environment and code structure of CORTEX-II have been designed to achieve a high degree of platform independence. ANSI C compatible compilers are used on both SUN workstations and PCs to allow algorithms to be developed and tested transparently in both simulated and real environments. This is accomplished through careful code structuring, as well as the use of the object-oriented technique of late binding. Late binding allows a software module to generate messages, such as navigation commands, without knowledge of the module(s) that receive the message. The binding between modules occurs at run-time, as opposed to the compile-time binding that occurs in standard function calls. Thus, an XView simulation module on the SUN can model navigation commands and robot kinematics, which are then used with no change in the C40 environment, where these messages would be captured by control software which would generate appropriate motor currents. Any direct function calls which require machine-specific software (such as display routines) are standardized and duplicated on all target platforms. This transparency between SUN-based simulation and C40 based run-time execution has provided to be a major convenience for developing applications.

5.7. Real-time Programming Environment

Software development on CORTEX-II has been done using 3L's Parallel C programming environment. Parallel C consists of Texas Instruments' ANSI compliant compiler, a set of library routines which provide support for various C40 features, a linker, a configurer, a debugger, a kernel for each of the C40s in the processor network and a server program which runs on a host PC. The library routines facilitate inter-processor communication across comm-ports, commands to peripherals via comm-ports and provide mechanisms which define and coordinate concurrent multi-threaded applications in addition to the standard C libraries. The configurer takes relocatable object modules and fixes them into a single application. This means that separate programs can be relocated to different processors without recompilation. A configuration file contains information which directs the configurer concerning which processor a module is to be placed on and which modules require communication links to one another. In this way, the connectivity scheme supporting the communication between modules is transparent to the modules themselves—they are solely concerned with their input ports and their output ports, not the sources and destinations of the data. The linking of ports between modules is left to the configurer. The PC-resident server loads the kernel and application code onto the C40s and provides the C40 kernels with run-time access to the PC's resources including the monitor and the file system. The PC is not required at run time, and our development system has an embedded PC merely to simplify loading the C40 kernel and application code.

5.8. Space-variant Image Processing Techniques

Image processing and pattern recognition algorithms are much more difficult in space-variant systems than in conventional imaging systems. There several reasons for this:

- *Lack of simple eight-connectivity.* Conventional TV rastors have a simple eight-neighborhood pixel structure in terms of which most conventional image processing algorithms are constructed. Space-variant frames, such as that based on the log-map, have neighborhood connectivity that is quite complex, and variable with position. This tends to "break" many algorithms for image processing, such as connected components, convolution, etc. Recently, we have developed a generic solution which allows image processing to take place on a pixel architecture with arbitrary connectivity. In this algorithm, the "connectivity graph" of the sensor is pre-computed, and all image processing primitives are defined in terms of this graph, rather than the implicit "Manhattan metric" that is usual. We have found that the connectivity graph algorithm, which has the same asymptotic complexity as conventional image processing, is capable of generalizing most common image processing algorithms, with the exception of the Fourier transform, to arbitrary pixel topologies, including random connectivity. This algorithm is described and demonstrated in Wallace et al. (1994).

- *Lack of shift invariant processing.* Space-variant vision systems based on the log-mapping have potentially useful size and rotation invariance properties (see Schwartz, 1980), but, by their very definition, these architectures greatly complicate shift-invariance. Image features change size and shape when the image is shifted across the field of a space-variant sensor. On a conventional TV rastor, shift invariance is easy to achieve, via Fourier techniques, auto-correlation, etc. But space-variant architectures, by definition, prevent the use of these techniques. Recently, however, we have found a solution to this problem. We have been able to define space-variant kernels, which, when convolved with the image, are fully shift, size, and rotation invariant. These kernels are convolved with the small space-variant image, and so the convolution can be done extremely quickly. In effect, this approach (Bonmasser & Schwartz, 1994) generalizes the conventional Fourier transform to a space-variant image structure.[13] We believe that this methodology will prove to be of fundamental importance to image processing and machine vision on space-variant domains, since it provides the capability for shift, size, and rotation invariant template matching, yet it utilizes the space-complexity advantages of the small log-map images, and so can be performed extremely quickly on hardware such as the C40.

6. SUMMARY

This paper has reviewed the theoretical basis for interest in space-variant active vision, has provided a general terminology for classifying the range of architectures that are generic to machine vision, and has outlined the specific practical problems associated with the space-variant active vision architecture. Solutions to a number of these problems have been provided by discussion of two recently constructed SVAV systems (CORTEX-I and CORTEX-II). The most important points made in this paper are as follows:

Commodity robotics. As reviewed in this paper, and derived in detail elsewhere (Rojer & Schwartz, 1990), the space-complexity of the log-polar map structure of primate vision is extremely favorable for problems in which wide-field and high-resolution must be jointly satisfied at the same time as minimum size, weight and cost. SVAV architecture provides an asymptotic space-complexity that is logarithmic in the ratio of field size to angular resolution, while the conventional space-invariant sensing architecture is quadratic in this figure of merit. The parameters of human vision suggest that up four orders of magnitude of advantage may be offered by this strategy, and this is almost certainly the reason that SVAV architectures are the only architecture represented in the higher vertebrate visual systems. The correlate of these observations is that a radical reduction in the size, cost, and weight of a computer vision system of a given field size/ resolution performance is possible via the SVAV architecture. If this resolution can match the 2–4 orders of magnitude that are theoretically possible, then it is to be expected that a new niche for machine vision will be opened, a form of 'commodity-robotics" in which widespread application of the formally restrictive techniques of computer vision may come to pass.

Biological and computer vision. The previous statement makes an assertion of a direct relationship between an anatomically and physiologically valid observation [log-polar architecture of primate vision (Schwartz, 1994)] and a practical application in machine vision. To date, this may provide the most firm connection between measurable data in the anatomy and physiology of biological vision and practical applications in computer vision.

Actuation. One major aspect of active vision in general, and SVAV in particular, is the importance of actuation. This may seem to be an obvious assertion, but experience dictates that the difficulties in finding appropriate actuation strategies are only fully appreciated by those who have tried to do so! There is an extreme shortage of viable actuation schemes which are simultaneously fast, cheap, small, light, and accurate. The final performance of an SVAV system is entirely dependent on the actuation scheme, and so, actuation rises to a prominent role in a field in which exotic CPU hardware and algorithm development tend to draw the most attention. In order to reinforce this point one more time, it is useful to consider the following quote from the Bureau of Naval Personnel Manual on Basic Optics and Optical Instruments, which makes a similar point, in a very grounded form, while justifying the study of glass to those interested in pursuing an optics specialization:

[13] Readers familiar with the use of the Mellin–Fourier transform should note that the Mellin transform is not space-variant, since it makes use of a log-polar mapping in frequency space, not image coordinates! Hence, the Mellin transform provides a "fovea" in frequency space, has no foveal structure in image space, and, moreover, requires computation in the full N^2 pixel space of a rank N image.

> "Without glass, there would be NO optical instruments, and no optical men in the Navy. . . [(Personnel, 1969), page 10]."

Similarly, without actuators, there is no active vision, and the invention of good actuation schemes, which motivated the development of the spherical pointing motor described earlier in this paper, will no doubt be a continuing motif in this field.

REFERENCES

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, **2**, 333–356.

Ballard, D. (1990). Animate vision. Technical Report 329, University of Rochester, Department of Computer Science.

Baloch, A. A., & Waxman, A. M. (1991). Visual learning: adaptive expections, and behavioural conditioning of the mobile robot mavin. *Neural Networks*, **4**, 271–302.

Bederson, B., Wallace, R. S., & Schwartz, E. L. (1992). A miniaturized active vision system. In *11th IAPR International Conference on Pattern Recognition*, Volume B of *Specialty Conference on Pattern Recognition Hardware Architecture* (pp. 58–62). The Hague, The Netherlands.

Bederson, B., Wallace, R., & Schwartz, E. (1994a). A miniature pan-tilt actuator: the spherical point motor. *IEEE Transactions on Robotics and Automation*, **10a**, 298–308.

Bederson, B., Wallace, R., & Schwartz, E. (1994b). Two miniature pan-tilt devices. *IEEE International Conference on Robotics and Automation*.

Bonmassar, G., & Schwartz, E. (1994). Geometric invariance in space-variant vision. In *ICPR Proceedings*, ICPR-12. International Conference on Pattern Recognition.

Burt, P. J. (1988). Algorithms and architectures for smart sensing. *Proceedings of DARPA Image Understanding Workshop* (pp. 139–153).

Clark, J. J., & Ferrier, N. J. (1988). Modal control of an attentive vision system. *Second International Conference on Computer Vision* (p. 514).

Engel, G., Greve, D., Lubin, J., & Schwartz, E. (1994). Space-variant active vision and visually guided robotics: Design and construction of a high-performance miniature vehicle. In *ICPR Proceedings, ICPR-12*. International Conference on Pattern Recognition.

Fiala, J., Lumia, R., Roberts, K., & Wavering, A. (1994). Triclops: A tool for studying active vision. *International Journal of Computer Vision*, **12**, 231–250.

Greve, D. N. (1995). Instrumentation and control of a miniature active vision system. PhD thesis, Boston University.

Greve, D. N., Engel, G., Lubin, J., & Schwartz, E. L. (1994). Actuator control of a miniature active vision system. In *ICPR Proceedings, ICPR-14*. International Conference on Pattern Recognition.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore, (Ed.), *Vision: coding and efficiency* (pp. 411–422). Cambridge: Cambridge University Press.

Ong, P. W., Bederson, B., Wallace, R., & Schwartz, E. (1992). Space variant optical character recognition. *11th IAPR International Conference on Pattern Recognition*, D:504–508.

Personnel, B. N. (1969). *Basic Optical Instruments*. New York: Dover.

Rojer, A. S., & Schwartz, E. L. (1990). Design considerations for a space-variant visual sensor with complex-logarithmic geometry. *10th International Conference on Pattern Recognition* (Vol. 2, pp. 278–285).

Rojer, A., & Schwartz, E. L. (1992). A quotient space through transform for space variant visual attention. In G. Carpenter & S. Grossberg, (Eds.), *Neural networks for vision and image processing* (pp. 407–436). Cambridge, MA: MIT Press.

Sandini, G. and Dario, P. (1989). Active vision based on space-variant sensing. *International Symposium on Robotics Research*.

Sandini, G., Bosero, F., Bottino, F., & Ceccherini, A. (1989). The use of an anthropomorphic visual sensor for motion estimation and object tracking. *Proceeding OSA Topical Meeting on Image Understanding and Machine Vision*.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, **20**, 645–669.

Schwartz, E. L. (1994). Computational studies of the spatial architecture of primate visual cortex: Columns, maps, and protomaps. In A. Peters & K. Rocklund (Eds.), *Primary visual cortex in primates*, Vol. 10 of *Cerebral Cortex*. New York: Plenum Press.

Schwartz, E. L., Merker, B., Wolfson, E., & Shaw, A. (1988). Computational neuroscience: Applications of computer graphics and image processing to two and three dimensional modeling of the functional architecture of visual cortex. *IEEE Computer Graphics and Applications*, **8**(4), 13–18 (July).

Shostak, S. (1992). The ultimate motion imaging system: What- and when? *Advanced Imaging*, 39–41.

van der Spiegel, J., Kreider, F., Claiys, C., Debusschere, I., Sandini, G., Dario, P., Fantini, F., Belluti, P., & Soncini, G. (1989). A foveated retina-like sensor using ccd technology. In C. Mead & M. Ismail (Eds.), *Analog VLSI Implementations of Neural Networks*. Boston, MA: Kluwer.

Wallace, R., Ong, P.-W., Bederson, B., & Schwartz, E. (1994). Space variant image processing. *International Journal of Machine Vision*, **13**(1) (in press).

Weiman, C. F. R. (1988). Exponential sensor array geometry and simulation. *SPIE Conference on Digital and Optical Shape Representation and Pattern Recognition*.

Weiman, C. F. R. (1989). Tracking algorithms using log-polar mapped image coordinates. *SPIE Proceedings on Intelligent Robots and Computer Vision*, **VIII**, 1192.

Weiman, C. F. R. (1990). Video compression via log polar mapping. *SPIE Symposium on OE/Aerospace Sensing* (pp. 1–12).

Yeshurun, Y., & Schwartz, E. L. (1989). Shape description with a space-variant sensor: algorithms for scan-path, fusion and convergence over multiple scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 1217–1222.