T-PAM/11/11//29341

Shape Description with a Space-Variant Sensor: Algorithms for Scan-Path, Fusion, and Convergence Over Multiple Scans

Yehezkel Yeshurun
Eric L. Schwartz

Reprinted from
IEEE TRANSACTIONS ON PATTERN ANALYSIS
AND MACHINE INTELLIGENCE
Vol. 11, No. 11, November 1989

Shape Description with a Space-Variant Sensor: Algorithms for Scan-Path, Fusion, and Convergence Over Multiple Scans

YEHEZKEL YESHURUN AND ERIC L. SCHWARTZ

Abstract—**One of the ways by which early human vision is sharply distinguished from machine vision is the fact that the human visual representation is strongly space-variant and the human system builds up a representation of a scene through multiple fixations during scanning.**

In this paper, we discuss three algorithms related to the "blending" of a single scene from multiple frames acquired from a space-variant sensor.

1) Given a series of space-variant contour-based scenes with different "fixation points," we show how to fuse these into a single, multiscan view, which incorporates the information present in the individual scans.

2) We demonstrate an (attentional) algorithm which recursively examines the current knowledge of the scene in order to best choose the next fixation point, based on focusing attention in regions of maximum boundary curvature.

Manuscript received June 1, 1987; revised March 23, 1989. Recommended for acceptance by O. Faugeras. This work was supported by the AFOSR under Contract F 85-0235, the System Development Foundation, and the Nathan S. Kline Psychiatric Research Center.

Y. Yeshurun was with the Computational Neuroscience Laboratories, Department of Psychiatry, N.Y.U. Medical Center, New York. He is now with the Department of Computer Science, Tel Aviv University, Tel Aviv, Israel.

E. L. Schwartz is with the Computational Neuroscience Laboratories, Department of Psychiatry, New York University School of Medicine, New York, NY 10016 and also with the Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012.

IEEE Log Number 8929341.

3) We discuss a simple metric for evaluating "convergence" over scanpath. This may be used to quantify the performance of (2) above, i.e., to compare the performance of various "attentional" algorithms.

Finally, we discuss this work in light of both machine and biological vision.

Index Terms—**Active vision, eye scanning, saccade, scan path, space variant vision, visual cortex.**

Introduction

When we view a scene, we have the subjective impression that what we see is stable and constant, both in position and resolution. However, this impression is far from correct. If we try to read a newspaper that is slightly off center (see Fig. 1), we become aware that the very high resolution provided in the region of our fixation (foveal projection) falls off rapidly toward the edges of our field of vision. The fact that the human visual representation is strongly space-variant implies that the human system builds up a representation of a scene through multiple fixations during scanning.

The space-variant nature of the human visual system is well understood, at least to the level of primary visual cortex. The threshold for visual acuity, stereo acuity, motion, and other psychophysical quantities scale at least roughly as the inverse of distance from the fovea. There is general consensus [1], [9], [10] that the spatial representation of the visual field,[1] at the level of the primary visual cortex, is approximated by a complex logarithmic mapping [4]. Fig. 1 and Fig. 6 of this paper show natural scenes processed by this form of mapping function. We are thus in a position to provide realistic estimates of the nature of a specific space-variant imaging system: that of the human.

In the present paper, we discuss three algorithms related to the "blending" of a single scene from multiple frames acquired from a space-variant sensor. We used contour-based scenes, rather than gray-scale scenes, in order to focus attention on the problem of space variance, as opposed to segmentation. The following generic problems are raised by considering a space-variant system.

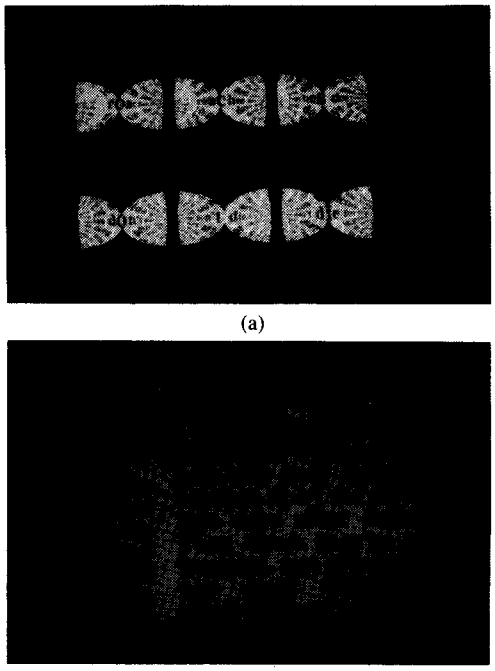
1) Given a series of space-variant contour-based scenes with different "fixation points," how might one fuse these into a single multiscan view which incorporates the information present in the individual scans?

2) How might one choose successive fixation points in order to rapidly gather shape-dependent data? Is there a simple attentional algorithm for contour-based scenes?

3) How could one quantify the rate of convergence of such a system as a function of the number of scans? What is the rate of convergence suggested by such a metric?[2]

In the present work, we do not address the classical issues of how the system (human or machine) is to obtain knowledge of its motor state (see [8]). Our intention here is to discuss the image processing problem of blending together multiple scans, obtained from a strongly space-variant sensor, and the problem of choosing a "scan path" which provides optimal information about the scene.

Assuming that a space-variant sensor similar to a human retina were available, it would be necessary to consider some of the issues discussed in the present paper: how should one choose a series of fixation points for such a sensor, how would one blend the successive frames, and how could one place a metric on the quality of this scanning process?



(a)

(b)

Fig. 1. (a) Simulates six successive scans of a newspaper, using a cortical map function derived from primate data [5], a reading distance of about 20 cm, and about 1.5° of visual field on each side of the fixation point. Each of the small "bow ties" represents the cortical "image" of a section of newspaper print. Thus, the first frame is fixated on the letter "o" in the word "roaches." There are two "bow ties" representing the left and right visual fields. The newspaper is then scanned, and the corresponding cortical "images" are presented in the figure. Note the strong space variance, even for the central few degrees of visual field. (b) Shows these six scans projected back to the visual field, and "fused" into a single scene [14]. The region of text scanned, which read "roaches don't die..," and to some extent the lines above and below this line, are seen clearly, but there is a rapid loss of detail in the text regions which are not close to the scanned text. Fig. 6 of this paper shows a wide-angle simulation of the human visual field and cortical image.

The Space-Variant Image and Boundary-Angle Function

We define the resolution at the point v of an image as the function $R_p(v)$ where p is the spatial location of a fixation point and R is a monotonic nonincreasing function of $|v - p|$. This is to say that R is proportional to the reciprocal of the minimal distinguishable distance (i.e., visual acuity). In the current context, the exact specification of R is not crucial; any R having the mentioned attribute can be used. The following discussion uses a function of the form $c/|v - p|$ for $v \neq p$ where c is a constant.

This definition might be applied to any gray-scale image (see Fig. 1). In other work, we address some of the difficult issues which arise when using gray-scale images [14]. In the current application, we consider only contour-based images in order to avoid dealing with issues such as segmentation. This situation can arise either naturally, when a scene is two-dimensional and consists only of contours, or artificially, after an edge-detection mechanism has been applied to an image of a complex three-dimensional scene (segmentation).

Boundary Contour Descriptor

In applications in which a one-dimensional representation of contours is desired, it is customary to use the boundary-angle func-

[1] In this paper, we do not discuss the detailed spatial architecture of primary visual cortex, which would include details such as ocular dominance columns, orientation columns, etc. We are only concerned here with the first-order topographic structure of the human visual system as a model for space-variant machine vision systems.

[2] In addition to these purely computational issues, the human system has also needed to: 1) evolve systems of accurate motor control [7], and 2) provide information to the organism about the current motor state (i.e., direction of gaze). This aspect of the problem has been much discussed under the terms proprioceptive perception, efference copy, corollary discharge, etc. [8].

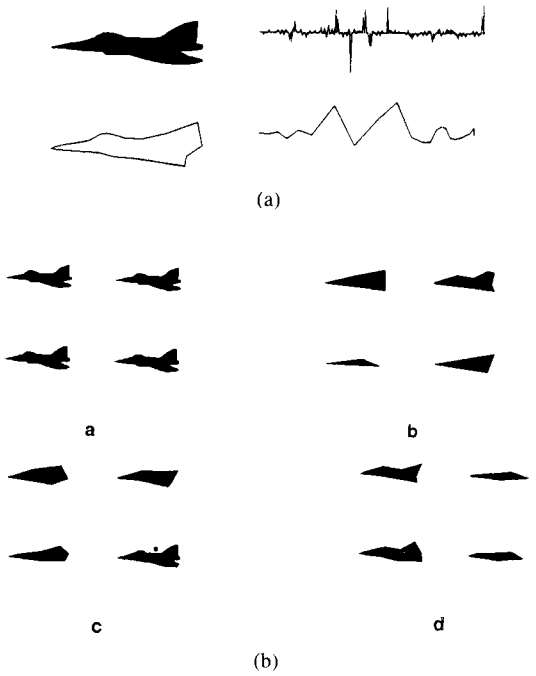


Fig. 2. (a) Images (left) and their boundary-angle functions (right). Top: the original contour (black silhouette) and its boundary-angle function. Bottom: the image as it is "viewed" from the fixation point (indicated by a star), with space-variant resolution. The tail of the airplane, being fairly far from the fixation point, is described very roughly. Therefore, the boundary-angle function bears only a rough resemblance to the original function. (b) A scene consisting of several planes, silhouettes (*a*), as it is "received" from different fixation points (*b*–*d*). The fixation points are depicted by an asterisk. The original airplane silhouette consists of 243 points, and the space-variant silhouettes average five points (for the less detailed ones) to 40 points (for the highly detailed).

tion $\theta(l)$, which is the angle of the tangent to the contour as a function of the arc-length l. In the current application, since we have discrete points connected by line segments (i.e., polygons), we use the representation $\Theta(l)$, which is the difference between two consecutive angles of the polygon. This one-dimensional representation of contours is most useful in shape-recognition tasks where it is further processed by a Fourier transform to yield the Fourier descriptors (FD's) of the contour [3]. There are also some indications that the FD of a shape might be useful as a shape descriptor in physiological studies of the primate visual system [6].

We apply spatial-variant resolution both to the image of the contour in the xy plane and to the boundary-angle $\Theta(l)$ representation of it, as explained below (see also Fig. 2).

1) The original contour is represented by line segments between the points $\{U_i, i = 1, k\}$. We assume that the distance between these points represents the highest possible resolution of the "viewer."

2) A new contour is defined by a fixation point: given a fixation point p and a contour point U_i, the value of $R_p(U_i)$ determines the next point U_j (i.e., by looking for the next point whose distance is at least $1/R$). Thus, starting at U_0, this procedure yields a contour whose points are a subset of the original points.

3) The boundary-angle function of the new contour $\Theta_p(U_i)$, $i \in \{1, k\}$ is obtained. To allow reconstruction of the original image, we also keep the resolution value $R_p(U_i)$ for each U_i.

In the xy plane, variable resolution produces a detailed image near the fixation point and a "blurred" image away from the fixation point. In the boundary-angle representation, the neighborhood of the fixation point is properly described, while other areas retain only smoothed, low-frequency details. The parameters used in this work yield a ratio of 1 : 10 between the full resolution image and a single space-variant view, which is in good agreement with the functional form of human visual acuity.[3]

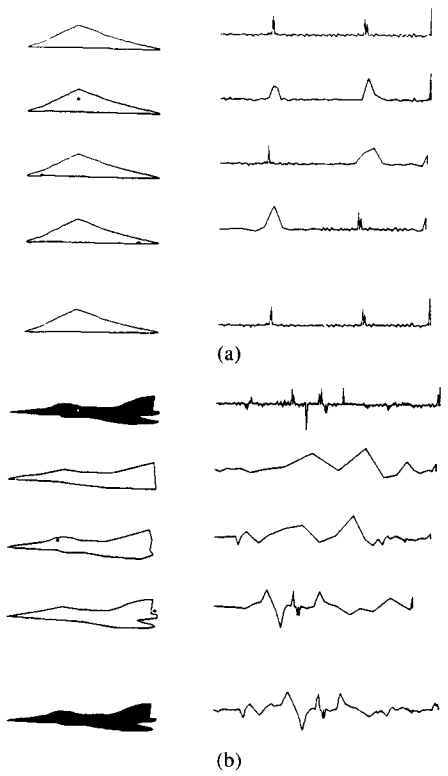


Fig. 3. (a) View of a triangle from three fixation points. The contour of the original triangle (top) is seen from three fixation points, each in the neighborhood of a particular vertex. These views are indicated by the corresponding boundary-angle functions. For each fixation point, only the closest vertex and its neighborhood are detailed, while the other vertices are approximated roughly. The reconstructed boundary-angle function (bottom) consists of the "best" contribution from each space-variant view. (b) A silhouette of an airplane, viewed from three fixation points, selected (by hand) because they are near areas containing many details. Details as in (a).

Blending Boundary-Angle Functions and Images

For a given fixation point, there exists a corresponding representation of the original contour. Several fixation points $\{p = p_1 \cdots p_n\}$ produce different representations of the same contour. This situation is shown in Fig. 3, in which images are viewed from several different points. Although the boundary-angle function $\Theta_p(U_i)$ is quite detailed near the corresponding fixation point, it just roughly approximates the original boundary-angle function in all the other areas.

Because resolution depends only on the distance between a given point and the fixation point, and because the most detailed boundary functions (or images) are obtained for high-resolution areas, an appropriate blending scheme should use the "best" of each view. The only information the blending scheme needs is the resolution associated with each point in the subcontour, which is kept when the subcontour is calculated. Thus, the reconstructed boundary-angle function is

$$\Theta^*(U_l) = \Theta_j(U_l)$$

[3]One recent estimate of primate magnification factor [1] suggests that there is a 10:1 decrease in spatial resolution of a stimulus between the fovea and 5° of eccentricity. This is a reasonable "viewing aperture" for shape perception. Note that a 10 : 1 (linear) change corresponds to a 100 : 1 area change, and that this area change is a more relevant index of "data compression."

such that

$$R_j(U_l) = \max_{p=p_1 \cdots p_n} \{R_p(U_l)\}.$$

The reconstructed function $\Theta^*(l)$ is an approximation to the original $\Theta(l)$. This approximation depends on the number of fixation points and their location. A more elaborate blending scheme might also depend on the "scan path" or sequence of fixation points humans select when viewing a given scene [12].

Choice of Scan Path: An "Attentional" Algorithm

Although early vision and artificial intelligence have received a great deal of attention recently, a great intermediate area exists which has received little study in this context, and that is the subject of "attention" itself. A single fixation provides partial information about a scene. Assuming that a unified representation of the scene can be extracted from successive scan, we must address the problem of locating the fixation points in such a way as to provide maximal information to the imaging system. This represents an ill-defined problem, as difficult issues relating to context and goal direction are implied by it. However, little advantage can be gained from a space-variant system without providing an attentional algorithm. In the following, we will discuss a simple candidate for attentional choice of successive fixation points.

In psychophysical contexts, the nature of visual scanning has been extensively explored (e.g., [2]). In general, fixation points tend to cluster around sharp edges, ends of lines, and locations where some "unpredictable" change takes place. Most existing research considers only the question of the location of the fixation points. Noton and Stark [12] have addressed the issue of choosing fixation points. They termed the temporal order of fixation points the "scan path," and found it to be consistent within a given subject and a given scene, but no further characteristics have been specified.

In our case, the scene consists of contours. The curvature of the contours is very likely to be a prime fixation-point "attractor" since large curvature represents rapid rate of change of boundary orientation. We can represent the curvature in terms of a boundary-angle function, indicating areas of high curvature by corresponding peaks in the function. A simple form of attentional algorithm, then, consists of the following steps.

1) Choose (randomly or by any method) an initial fixation point.
2) Calculate the boundary-angle function according to the current fixation point.
3) Select the next fixation point according to the maximum of the boundary-angle function $\Theta_p(U_i)$.
4) Keep the boundary angle function and the corresponding resolution values. Keep a reference point in the current fixation that will be associated with a point in the next fixation.
5) Blend the views and the boundary angle functions to yield a single view/function.
6) Go to step 2) until "convergence" (see below).

Such a procedure is shown in Fig. 4. The fixation points in this figure seem plausible in comparison to the points that one would likely select without using the algorithm. However, the algorithm has one drawback. In cases where several high values of the boundary-angle function cluster together, the algorithm picks several fixation points at almost the same place. Because the scans obtained from adjacent fixation points do not differ much, and because the foveal area can cover several points of high curvature, this clustering of points is redundant.

In order to remove the redundancy, we modify the algorithm [in step 3)] by considering $\Theta(U_i)\,W(U_i)$ instead of $\Theta(U_i)$. The weight function $W(U_i)$ can be used to enhance (or mask) selected features. If W is chosen such that it equals 1 everywhere except for a neighborhood of the fixation point where it vanishes, the redundancy problem is solved. In other words, after a fixation point is selected, the relevant foveal area (i.e., the area immediately surrounding the fixation point where the high resolution still holds) is not counted when the algorithm searches for the next higher value. Fig. 4(b) shows the results of this approach.

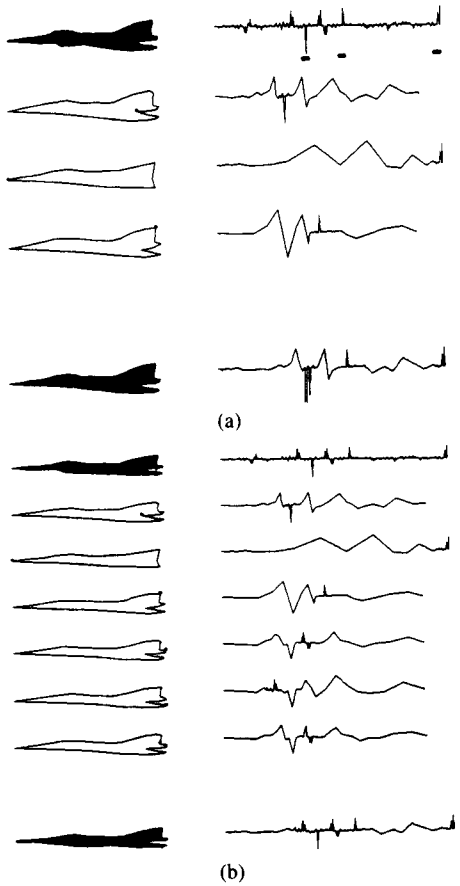


Fig. 4. (a) Images (left) and the corresponding boundary-angle functions (right). The top row shows the original image and function; the next three rows represent three fixation points (denoted by small stars on the images), and the bottom row shows the integrated image and function. The fixation points, which are selected automatically, are the spatial locations that correspond to the three largest values of the original boundary-angle function (denoted by bars under the function). (b) Results of the modified algorithm. The fixation points are chosen by the maximum of $\Theta(U_i)/R(U_i)$.

One might also select W to be $1/R$, thus emphasizing "remote" features rather than "close" ones. Finally, W might contain some random fluctuations in order to avoid the possibility of being "trapped" between two features.

The algorithm needs a reference point that is shared between each two successive fixations: this is necessary when the views or the boundary-angle functions are "tailored" together.

Convergence and Norms

Because our figures consist of simple contour drawings, it is easy to define a norm that compares composite space-variant scenes after n scans with the original high-resolution scene. A reasonable choice for this norm is a least-squares measure of the two boundary-angle functions. Thus, let Δ_n represent the difference between the full-resolution scene and the composite scene after the incorporation of the nth fixation point: $\Delta_n = |U - C_n|$.

Using this norm, it is possible to define the convergence rate as a function of the scan path. Thus, for a sequence of fixation points $p_1, p_2, \cdots, p_n$, we define the rate of convergence for the scan path at point n as $\Delta_n - \Delta_{n-1}$. This method is suitable for the purpose of the algorithm's evaluation or for calibration when we have access to the full-resolution contour. However, in a "real-time" situation (i.e., in robotic vision), the full-resolution image is not necessarily available. Thus, we can define Δ_n as $|C_n - C_{n-1}|$, and base the "convergence" decision on it (see Fig. 5). If one thinks of n as a time variable, then this measure indicates the "rate" of error reduction.

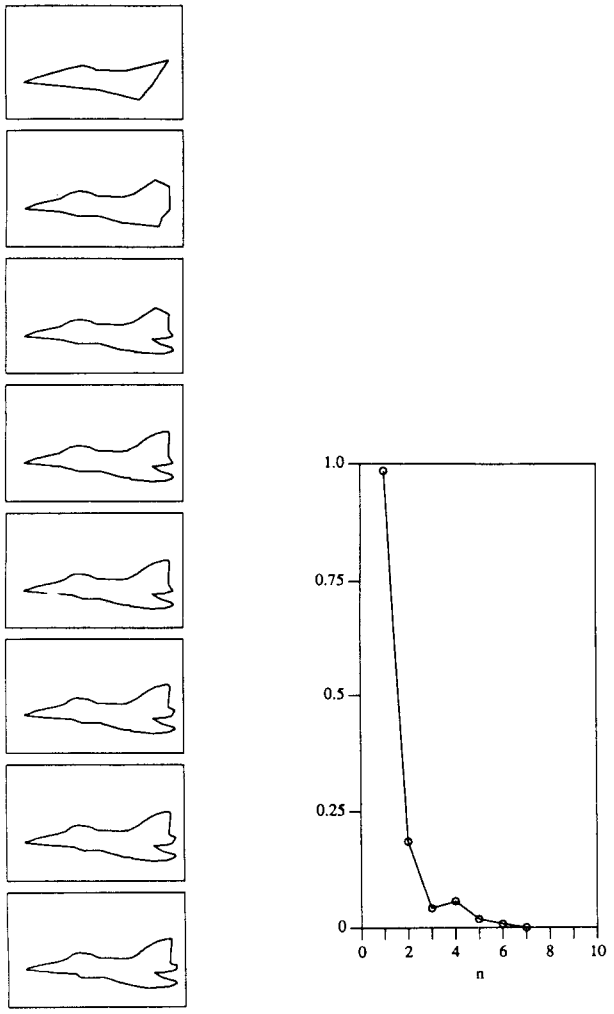


Fig. 5. Converging rate of the algorithm, as depicted by the difference Δ_n between successive blended figures. Left: blended figures after 1, 2, 3, $\cdots$, 8 fixation points. Right: Δ_n versus number of fixation points.

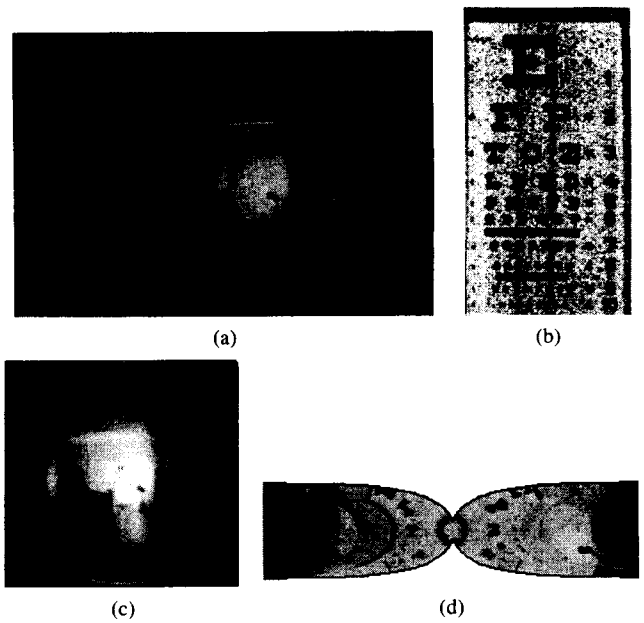


Fig. 6. (a) shows a wide-angle fish-eye view of a scene in the hall of our laboratory. A ladder is to the right, an eye chart is in the very center of the frame (almost invisible). The original version of this scene was digitized to an effective resolution of 16 000 × 16 000 pixels by a polar-coordinate mosaic technique. A "blowup" of the central region of this original frame is shown in (b). This is an eye chart, and the distance to the chart was 20 ft. In the original, line 7 of the chart could be easily read, indicating an effective "acuity" of 20/30 or about 1.5 min of arc. The purpose of this work was to simulate a wide-angle scene (about 100°), roughly comparable to human vision, at human visual acuity. (c) shows this scene, blurred by a space-variant filter which is modeled after human visual acuity. (d) shows the image of (a), modeled in terms of a complex-logarithmic model [4] of human visual cortex. The eye chart occupies almost half of the surface of visual cortex, although it occupies a tiny fraction of the original scene. The ladder and the windows of the original are compressed to almost the same size as the centrally fixated letters of the eye chart. This illustrates the tremendous space-variant compression of human vision. Variations in linear size of about $100^2:1$ (10^4 in solid angle) occur from the center to the periphery of the human visual system.

Thus, one algorithm for adding scan paths might be based on the addition of a new point which, among all the possible fixation points, maximizes the above "rate" of convergence. Conversely, the addition of new points becomes unnecessary when no points can be found that significantly improve the rate of convergence. The algorithm we propose rapidly converges: it is monotonic in the sense that only "better" resolution points are introduced, and it is bounded by the original set of points which constitutes the object. Fig. 5 shows an example of an aircraft silhouette which is scanned by this algorithm, with a plot of convergence based on the latter method described above. It is clear that there is rapid convergence to an accurate representation of the boundary of the figure. It is interesting to note that Noton and Stark [12] report that humans typically view scenes with perhaps three–eight scans; our algorithm also converges quite rapidly, in this case in which parameters of space variance derived from human vision have been used.

In more general cases, however, the choice of a norm is likely to be quite difficult. In the general case, both the attentional algorithm and the norm used to evaluate its success would likely be dependent on past experience, the goal-directed state of the imaging entity, and the full context of the current task. In lieu of engaging in this full-blown algorithmic study of visual attention, we propose that the simple curvature-based norm and scanning algorithm outlined above provides an initial step in the direction of understanding visual attention, and is one which is optimal in those situations in which a value-neutral estimate of boundary curvature is the desired information.

Implication of Space-Variant Image Processing to Gray-Level Images

Although we address mainly contour-based images in this work, it might be of interest to point out its application to gray-level images, especially from the aspect of "data compression."

The human visual field subtends more than 100 × 100° [13], with a maximum resolution of about 1 min of arc (foveal). Using a space-invariant sensor (e.g., conventional CCD camera), one would have to resolve 6000 × 6000 samples (1 min of arc × 100° in each direction). In order to achieve this performance, one would have to sample at two–three times this resolution in each dimension. An image of 16 000 × 16 000 would provide this performance, but would extend close to the gigapixel range in size.

We have experimentally demonstrated this estimate by digitizing[4] a conventional eye chart, at a distance of 20 ft, using a wide-angle (fisheye) lens, which recorded from about 80° of field. Fig. 6 shows the "full scene" and a highly magnified detail of the eye chart at the center. We continued to magnify the scene until the 20/20 line of the eye chart was visible (indicating a resolution of about 1 min/arc). We calculate that this occurred at an effective sampling resolution of 16 000 × 16 000 pixels.

Although both of the previous estimates are ad hoc, they agree well enough to suggest that the effective resolution of a single scan

[4]We used a conventional NTSC frame grabber, at 480 × 525 resolution, together with a polar coordinate mosaic technique [11] to produce this simulation.

of the human system is equivalent, were it recorded by a space-invariant system, to a 1/4 gigapixel image. Now, this estimate of 1/4 gigapixel is based on the use of a constant resolution system, which extended over 100 × 100° at full visual acuity. In fact, we simulated the logarithmic structure of the human visual system, and our simulated image occupied only about 16 000 pixels (see Fig. 6). Naturally, we only obtained high resolution over a small "foveal" representation with this simulation; in order to use this approach effectively, multiple scans would need to be performed. However, with a relative data compression of about 16 000 : 1, we can afford to perform the scanning process over a number of fixation points. Even 16 successive fixations would yield an effective 1000 : 1 data compression relative to a constant resolution system, provided that one obtained a satisfactory representation of the image regions of interest.

Summary

Space-variant imaging has been little explored in the context of machine vision, but is a major area of interest in the context of biological vision. Space-variant imaging provides a number of advantages, and difficulties, with respect to conventional space-invariant systems. One advantage is that very large fields of view can be covered, and very high resolution can also be provided. This leads to a form of image data compression which can be extremely large. However, a number of algorithmic difficulties are introduced by considering strongly space-variant systems. Attentional algorithms are required to make effective use of the small high-resolution "fovea," while other algorithms are required to "fuse" successive space-variant scans.

In the present paper, we have provided preliminary solutions to each of these issues. Using our algorithms, we obtain satisfactory convergence, for reasonable parameters of space variance derived from human vision, over a small number of scans (perhaps three-five scans).

The possibility that space-variant sensors (e.g., CCD's) may become available for application in machine and robotic vision should provide some motivation to begin studying the issues which such a sensor would provide. Perhaps the possibility that some of the high performance of the human visual system derives from its use of a space-variant architecture may provide some impetus to develop such a sensor.

References

[1] B. M. Dow, A. Z. Snyder, R. G. Vautin, and R. Bauer, "Magnification factor and receptive field size in foveal striate cortex of monkey," *Exp. Brain Res.*, vol. 44, pp. 213–228, 1981.

[2] K. Rayner, "Eye movement in reading and information processing," *Psychol. Bull.*, vol. 85, pp. 618–660, 1978.

[3] C. W. Richard and H. Hemami, "Identification of 3D objects using Fourier descriptors of the boundary curve," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, no. 4, pp. 371–378, 1974.

[4] E. L. Schwartz, "Computational anatomy and functional architecture of striate cortex: A spatial-mapping approach to perceptual coding," *Vision Res.*, vol. 20, pp. 645–670, 1980.

[5] ——, "On the mathematical structure of the retinotopic mapping of primate striate cortex," *Science*, vol. 227, p. 1066, 1985.

[6] E. L. Schwartz, R. Desimone, T. Albright, and C. G. Gross, "Shape recognition and inferior temporal neurons," *Proc. Nat. Acad. Sci.*, vol. 80, pp. 5776–5778, 1983.

[7] D. L. Sparks and J. D. Porter, "The spatial localization of saccade targets II: Activity of superior colliculus neurons preceding compensatory saccades," *J. Neurophysiol.*, vol. 49, pp. 64–74, 1983.

[8] R. W. Sperry, *J. Comp. Physiol.*, vol. 43, pp. 482–489, 1950.

[9] R. B. Tootell, M. S. Silverman, E. Switkes, and R. deValois, "Deoxyglucose, retinotopic mapping and the complex log model in striate cortex," *Science*, vol. 227, p. 1066, 1985.

[10] D. C. Van Essen, W. T. Newsome, and J. H. R. Maunsell, "The visual representation in striate cortex of the macaque monkey: Asymetries, anisotropies, and individual variability," *Vision Res.*, vol. 24, pp. 429–448, 1984.

[11] E. Wolfson and E. L. Schwartz, "Space-variant image-processing II: A truncated-pyramid algorithm for giga-pixel image-warping," *Comput. Neuro. Tech. Rep.*, vol. CNS-TR-23-86, NYU Med. Cen./Computational Neurosci. Lab., 1986.

[12] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision Res.*, vol. 11, pp. 929–942, 1971.

[13] M. Levine, *Vision in Man and Machine*. New York: McGraw-Hill, 1985.

[14] E. Wolfson, Y. Yeshurun, and E. L. Schwartz, "Space-variant image-processing II: Image-blending of multi-fixation logarithmic views," *Comput. Neuro. Tech. Rep.*, vol. CNS-TR-10-86, NYU Med. Cen./Computational Neurosci. Lab., 1986.