CHAPTER 11

Multiple Scale Analysis

11.1. Introduction

In previous chapters it has been shown how it is possible to derive a great deal of high fidelity data from images by spatially and temporally interactive processing based on an interpretation of human vision. One fact which *must*, however, be acknowledged, is that the *incoming* information content in a scene, as imaged by any optics and sensed by any sampling system, must be dependent on the distance of the scene detail from the imaging optics, the magnifying power of the optics (i.e. its focal length) and/or the physical size of the sampling cells in the sensing system. Consider for a moment what appears to happen with human vision. We are provided with an essentially fixed focal length imaging optics (and therefore a fixed magnification of the incoming image on the retina). There is thus a strict *minimum* size of detail, defined by the physical imaging properties and the sampling theorem [e.g. 11.1], which we can intelligently interpret. Above this minimum, however, there is a wide range of sizes of image detail over which we appear to be able to interpret readily, immediately and consistently. For instance, a head is instantly seen as a head, whether it is so small that it is a minimally resolved blob or it occupies several degrees of the visual field. This capability appears to exist not only independently of size, but also independently, as well as in parallel with, interpretation of any internal resolved details which may be present. One might speculate that such a facility is *necessary* for a versatile image processor, bearing in mind the uncertainties of presented image size in many visual environments. It is therefore worth addressing the question "How is this size independent performance achieved?".

11.2. Some relevant observations on human visual function

For the moment still considering human vision, it has been acknowledged that at least part of its impressive capability may come about as a result of the organisation of the retina. The eye has a fovea for high resolution vision, surrounded by a periphery where both image blur and sampling interval increase progressively and dramatically (e.g. [11.2] and Chapter 2). Note that the change is *progressive*, rather than a relatively discrete annular region of much lower resolution as has been assumed by a number of workers. These progressive changes automatically provide *some* facility for scale independent form perception –

obviously if one looks directly at an object, then, as the object grows in size, so its main profile moves into the peripheral regions of the eye and thereby is perceived at a progressively lower resolution. However, it is unlikely that the inhomogeneity can explain the *entire* size independency and, anyway, such inhomogeneity is not readily compatible with conventional computing environments. *Some* considerations have nevertheless been given to such possible processing arrangements both by us and by others. These will be discussed a little in Chapter 18.

Yet again, the eye is known to have around four scales of spatially interactive receptive field operating in parallel at any one location (see Chapter 2). It has therefore been suggested by some workers that this leads to four channels individually maximally responsive to four sizes of object. Whilst this certainly must be true, in principle, equally it seems unlikely that this is the entire explanation of the observed size independency. In particular, in human vision the four channels are *recombined* into one channel at the optic nerve (see Chapter 2), so they can hardly be called independent at the perceptual level.

Since all four basic channels *are* recombined, one can, for any particular local image situation, only consider *one* predominant signal from them at any one instant of time. Also, for any combination of outputs from a series of channels of an excitatory/inhibitory centre/surround nature, it is necessary to obey a law that the strength of the signal from the centre of one channel is approximately equal to the signal from the surround of the next smaller one [11.3, 11.4], in order to avoid serious inter-channel noise problems. Accepting this, one must realise that, in general, if there is a strong profile or hot-spot signal, sensed by the highest resolution channel, any signals sensed by the local lower resolution channels will be relatively insignificant. Conversely, at locations away from strong profiles and hot spots, any more distributed first difference signals will be readily sensed. Exceptions to this rule are certain adjacent profile situations, such as those studied by Fiorentini and Maffei – [11.5, 11.6] – King-Smith and Kulikowski – [11.7] – and Westheimer – [11.8], where the human visual system shows evidence of the expected inter-channel interactions. See [11.9] for a thorough appraisal of these special situations. The foregoing is in no way in disagreement with work of such people as Bergen and Wilson [11.10 and 11.11], since, for the specific (and unnatural) case of one-dimensional sinusoidal patterns used by such workers, which have no sharp profiles (i.e. low frequency patterns), there will be no major stimulus for the high resolution channel and one of the others will certainly be predominent. However, in the real world we are usually mainly concerned with profiles, be they from object contours or from texture. We have isolated other functions than profile sensing for the three lower resolution channels when viewing natural scenes – that is, largely as progressive biasing functions to ensure that, for highly structured scenes, the local adaptation level at various parts of the retina is roughly optimised for efficient function of the highest resolution (or form) channel. This is bearing in mind the fact that, at a given adaptation level, the effective dynamic range of receptor response is only about two orders of magnitude of illuminance [11.12]. Hence it is considered that, for form information, it is effectively only *one* channel (the highest resolution channel) which transmits data

to the cortex. Thus we must still pose the problem of size independence (i.e. a head is immediately perceived as a head, whether one can see fine details within it or just the general profile) and additionally we must ask how global analysis can be achieved.

11.3. Some relevant practical observations concerning computer image processing

Let us now turn to some practicalities of multiple scale *computer* image processing. On the one hand there would appear to be a desirability for a multiple scale processing in order that, *at one scale or other*, there will be just sufficient data available to provide a clear outline of features of interest without internal clutter. This aspect of multiple scale processing also would provide a ready means of 'playing off' fidelity of information against noise suppression (a technique well illustrated in principle by Baker and Sullivan [11.3]). On the other hand I have already evidenced that, for local motion and stereo perception, in order to provide a high sensitivity, one must inevitably compromise (Chapter 5). One must accept a strictly finite (around 2 pixels) limit to the maximum displacement between a pair of frames which can be derived reliably from a single channel system. At the same time for *global* motion and stereo it is necessary to provide a facility for sensing *large* displacements, but with no overt need for high image fidelity. Thus, once again there is a need for multiple scales of 'perception'.

One possible solution to the requirement for multiple resolutions, which has been suggested by several authors, is to have a lot of image processors operating in parallel at different scales and with outputs interlinked (e.g. multiple channel DOG processors [11.3]). However, such multiple channel processors involve widescale, accurately weighted integration for the broader channels, and are thus cumbersome and slow or expensive. To reduce the computation to some extent, some workers have recently drawn attention to a useful property of Gaussians – that one Gaussian convolved with another Gaussian yields a third Gaussian. By such means they claim to be able to reduce the massive broad channel convolutions dramatically by repeatedly convolving with a smaller Gaussian. In practice, however, simple repetitive convolution with a fixed small Gaussian kernel, and retaining a fixed sampling interval, is *not* particularly satisfactory. The resultant standard deviation from convolution n times with a Gaussian of standard deviation σ is only $\sqrt{n} * \sigma$. This is hardly efficient for generation of widescale blur. Furthermore, when the resultant is a *very* blurred input to the subsequent spatially interactive processes, this input will be highly oversampled. Strictly speaking, for any given composite Gaussian spread there is an optimum sampling interval, this sampling interval being proportional to the standard deviation of the Gaussian (e.g. [11.13]). This is equivalent to saying that one may obtain all useful information from a scene related to a broad Gaussian by *scaling down* the input image and then processing with the original Gaussian (a massively reduced computation).

In the appropriate chapters, *for illustrative purposes only*, an ability to handle

large displacements of a global nature, by use of scaled down *input* images, where the true structure is hardly recognisable, has been already demonstrated (e.g. Chapters 9 and 10). This is equivalent to what *would* be the likely human visual processing, if the various retinal receptive fields were *not* recombined at the ganglion level. What of the possibilities for *serial* processing at a variety of scales?

It is believed that we have a simple solution to the multiple scale problem available to us in the fragmentary outputs of VISIVE as discussed in Chapters 3 and 4. If, in addition to the conventional fragmentary bar and edge arrays which we have built into VISIVE, we permit looping back to the partial second difference signals simulating the high resolution channels at the optic nerve, then one may loop around the bar and edge extraction routines, whilst progressively shrinking the array. If at each looping one includes a local weighted integration, together with a resampling at increased spacing (to maintain optimum blur versus sampling), then each new pixel contains partial second difference signals which are a weighted average of those in an original local grouping (in strength, position and orientation), whilst generating new, smaller hexagonal arrays. Processing these smaller matrices through the bar and edge routines yields secondary first and second difference profile data of reduced resolution, but containing all the form, position, motion and/or stereo disparity data for that reduced resolution. Thus, in very few iterations, such a looping generates, progressively, lower and lower resolution scene data down to, ultimately, one global output, all from the *one* channel! Since the processing is progressive, one may anticipate detection of a local low resolution hot-spot at one plane, followed by tracing back through previous planes (top down processing) to assess higher level local feature information. Local relationships between planes are very simple to arrange, because of the pipeline nature of the processing. This is seen as one form of simple practical solution to the conceptual scheme discussed by Granlund [11.14, 11.15]. It is also seen to have very considerable connections to concepts of quadtrees [e.g. 11.16, 11.17], octrees [e.g. 11.18] and septrees [11.19], and to the techniques employed in the split and merge form of region segmentation ([11.20] and Chapter 15). There is a particularly close association with the septrees recently published, since the progressive blocking on a **pseudo-*hexagonal*** matrix used in this concept, whilst maintaining a pixel centre reference, fits in very well with our concepts of multiple scale windowed analysis (Chapters 15 and 16). However, all the 'tree' techniques, and split and merge segmentation to a large extent, rely on progressive breaking down of an image into smaller and smaller areas of similar *grey level*. The 'progressive shrink' and window approaches, on the other hand, are driven by significant grey level differences, irrespective of actual local grey levels. For instance a very large, highly resolved target with lots of texture should be seen as a mass of texture at plane 1, and then progressively as a detailed outline, a poorly resolved outline, an unresolved hot-spot and a sub-threshold component of the general scene. At each and every level, data can be extracted on fragmentary orientation, centroid, local motion and/or stereo disparity. In the limit the final plane should contain measures of global 'busyness' of the scene (or entropy), mean global stereo disparity and, in the temporal domain, global optical flow.

The progressive coalescence for a one dimensional situation is shown conceptually for blurred input images in Fig. 11.1. Here we have an image consisting of two long bar features separated by several tens of pixels. Initial stages of such scale reduction serve to condense the information from the two bar features of finite width to two unresolved line features (stage 3). Further scale reductions then gradually cause the two line images to coalesce, in the end yielding a single unresolved line feature (stage 6). It should be noted that such progressive coalescence of image details will only occur correctly for the original image or for partial or complete second difference arrays. Attempts to carry out similar coalescence *directly* in *first* difference space will fail. Hence in order to arrive at a set of progressively scaled *first* difference arrays (as recommended for efficient form analysis in Chapter 4), as well as progressively scaled *second* difference arrays (as recommended for motion and stereo analysis – Chapters 9 and 10), it is necessary

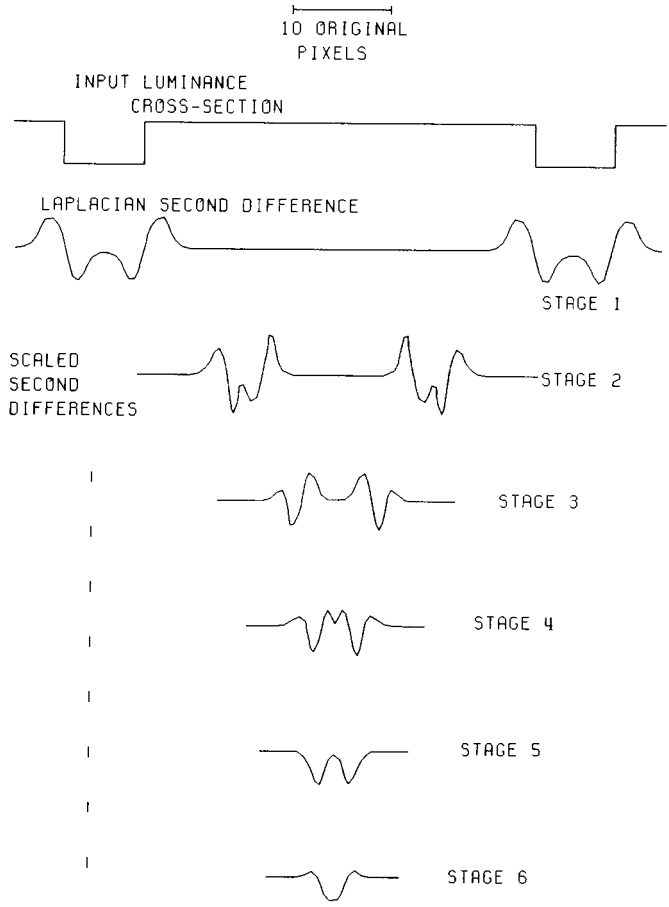

Fig. 11.1. Illustrating the effects of progressive blurring and scale reduction in second difference space for a pair of widely separated bar images.

to carry out the *practical* processing on the *positive* and *negative components* of second difference (Chapter 4). This fits in beautifully with conceptual *human* visual processes, since these individual parts of the second difference signal seem to be retained to a very high level of the visual processing tract (Chapter 2).

11.4. Concepts for efficient multiple scale analysis

Two important factors which are absolutely necessary for any robust, high fidelity image processor are:

(i) that there is a good balance between image blur and sampling [e.g. 11.13] and
(ii) that there is a correct balance maintained between noise contributions from various pixels when any spatially interactive processing is carried out.

If these factors are neglected then, although the processor might work well for simple, low noise images, it will be doomed to failure in some 'difficult' situations. A great deal of effort has gone into establishing the optimum balance in the present forms of VISIVE. The input optical spread to be sampled is of approximately Gaussian form and with a standard deviation of around 1.3 pixels (e.g. Chapters 3 and 4) (near to the optimum for information transfer by a sampling system [11.13]). This blur/sampling relationship is the key to a wide variety of simple data analyses – e.g. vernier position and vernier orientation (Chapters 3 and 4), motion sensing (Chapters 5 and 9), stereo disparity sensing (Chapter 10), corner cueing (Chapters 12 and 13) etc.. If we are to attempt to carry out progressive scaling operations, and expect to be able to perform similarly detailed analyses on the reduced scale images, it is thus essential that the blur/sampling relationship is roughly maintained as the scaling proceeds. This in turn implies some sort of progressive local weighted integration. Yet again, early versions of VISIVE, attempting to carry out simple bar and edge detection, were found to result in a serious imbalance when dealing with noisy images, even on the basic full scale outputs (unpublished studies). Great care was taken to find a series of local processes such as bar integrations which were so weighted that the composite result was adequately balanced for random noise. The forms of local functions were studied with both hexagonal and square matrix versions of VISIVE. The approximate functions for our preferred hexagonal matrix VISIVE appear to work well, but have not been extensively studied. They are as follows:

(i) Local weighted integration of the form shown in Fig. 11.2a. This, when compounded with the simple Laplacian operation, provides well-behaved interaction with following receptive fields, whilst operating throughout with integer weightings. It also provides a plausible agreement with electron microscopy evidence of both primary and secondary ring inhibition in primate retinae [11.21].
(ii) Three bar detectors at 60 degree intervals (primary axes), which are essentially 5×1 fields with weighting such as shown in Fig. 11.2b, plus three bar detectors at intermediate intervals (secondary axes) with form and weighting such as shown in Fig. 11.2c. Note particularly the *zero* weighting of the central cells.

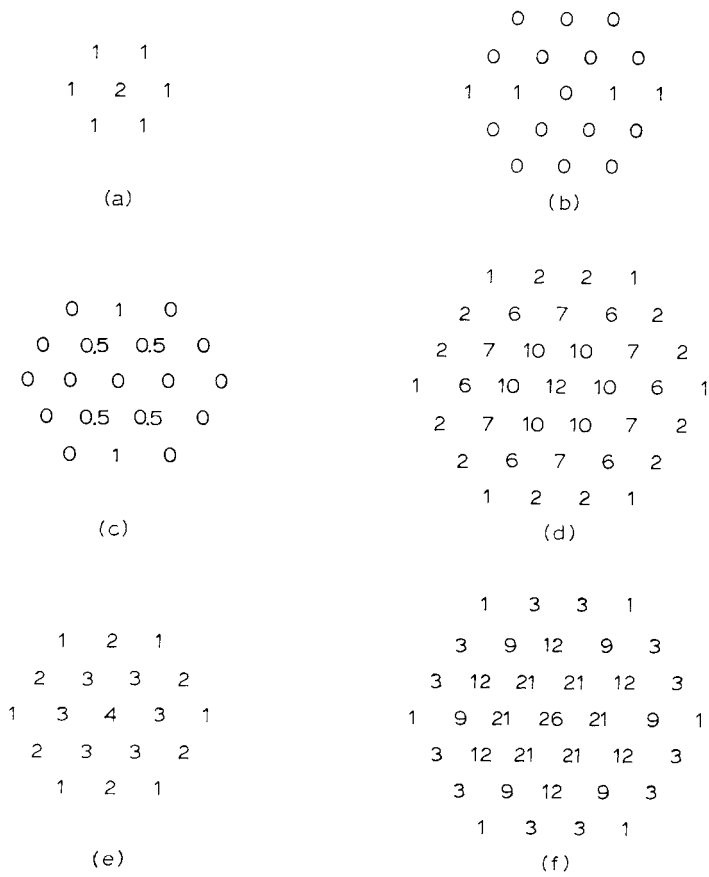

Fig. 11.2. Integer-weighted local neighbourhoods studied for progressive blurring and smoothing. a) Local weighted integration. b) Primary axis bar integrator. c) Secondary axis bar integrator. d) Combined effects of (a), (b) and (c). e) 3 ring 2-D weighted local integrator. f) 4 ring 2-D weighted local integrator.

This is necessary in order to avoid unfair noise contribution when considering the composite effect of six orientation-tuned outputs.

The effect of these seven operations, when compounded at the summed bar and summed edge arrays, is then as shown in Fig. 11.2d (note that this is still integer weighted). We then have, at the summed bar and summed edge arrays (from which major vernier motion and stereo extraction is carried out in present versions of VISIVE – Chapters 4, 5), noise balanced first and second difference data arrays which are already rather more blurred than the initial sampled image. It was considered, therefore, that this should be a good place from which to start a progressive scene contraction. What was proposed was that further local neighbourhood convolutions be carried out on the positive and negative parts of the summed bar outputs, using a convolution matrix which itself was noise balanced. This was necessary in order to increase the blur to a roughly Gaussian form of

approximately double the width of the input spread function. It was then proposed that this should be resampled every alternate pixel and line, to yield secondary hexagonal matrices of linear scale reduction of 2:1 (see Fig. 11.3). These new matrices would then be offered to the later stages of VISIVE as if they were the original positive and negative partial Laplacian second difference arrays. It should then be possible to compute, using standard VISIVE algorithms, all the vernier, motion, stereo disparity and feature data normally computed from VISIVE, but for the reduced scale image. In the process, new positive and negative partial summed bar arrays would be generated. These in turn could serve as the starting point for a further contraction, until such time as the residual arrays were too small to work with. In carrying out a series of 2:1 linear scale reductions it should be realised that the number of data points reduces by 75% on each pass and thus, from standard summation for geometrical progressions, the total data output after *all* physically realisable contractions would be less than 1/3 greater than that from the first pass alone. That is, the sum to infinity for a geometrical progression with ratio $0.25 = A/(1 - 0.25) = 1.33A$, where A is the data in the first process.

11.5. Practical implementation

The smallest local integer weighted neighbourhood operator which can be considered, for providing the necessary additional blur with noise balance in conjunction with the partial summed bar arrays, is the seven pixel unit as used for local integration after Laplacian operation in basic VISIVE (Fig. 11.2a). However, when this is convolved with the summed bar array, the total secondary blur after the Laplacian is grossly insufficient for the scaled down image. The next possible noise balanced, integer weighted local operator is of the form shown in Fig. 11.2e, which again is found to yield a composite secondary blur which is grossly insufficient. The idea of going to much larger local neighbourhood operations is computationally unattractive, so an alternative was sought. Now it is necessary, for

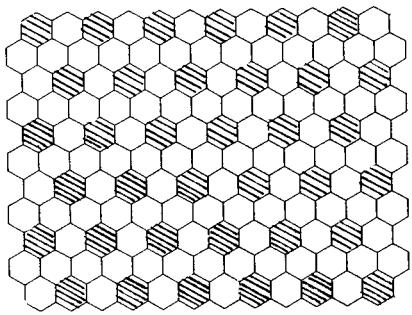

Fig. 11.3. Illustrating the resampling necessary for 2:1 scale reduction with a hexagonal matrix. ⊗ Pixels sampled for the reduced scale matrix. Note the stagger on alternate sampled lines (even numbered new lines staggered to the left for consistency with standard VISIVE processing conventions).

scaling, to sample every *second* pixel and every *second* line according to Fig. 11.3 (as discussed previously). The question may therefore be posed – "Can we resample *first* and *then* operate by a local neighbourhood?". This was investigated for the two local operators previously discussed. It was found that the seven cell operator *still* yielded a composite blur which was insufficient in the new sampling domain (a half-width of only about ± 1 pixel). However, the 19 cell operator yielded a composite as shown in Fig. 11.2f. This can be seen to have a half-width of approximately ± 1.5 pixels *in the resampled space*, which is almost ideal. It was therefore proposed that this local neighbourhood operator be used. Having resampled and carried out the convolutions, the resultant arrays were considered as replacements for the partial Laplacian second difference output arrays. As such, they were processed through VISIVE in exactly the same way as the original partial second difference data. This yielded a set of outputs of vernier position, vernier orientation, edge strength and local motion and/or stereo disparity perpendicular to local profile fragments for the reduced scale (lower resolution) scene, plus new partial summed bar arrays. These new partial summed bar arrays had the correct attributes to be treated in the same way as the original partial summed bar arrays. These yielded a further reduced scale input to VISIVE at the partial second difference level, a third set of fragmentary profile data and so on. In addition the data at each successive level were at a progressively improved signal/noise (due to standard noise suppression by averaging). Also, considering the temporal domain, the progressively reduced temporal resolution reflected from the reduced spatial resolution provided a facility for sensing progressively increasing motion bands.

An experimental extended version of the early, zero-crossing finding form of VISIVE incorporating the foregoing suggestions was set up some years ago. Initial runs were extremely promising. A typical sequence of outputs from that study at reducing resolution is shown as Fig. 11.4. [For an explanation of the method of plotting data used here, see Chapter 4, Section 4.4] It was intended at that time that a thorough assessment of this technique should be carried out in order to assess the differences, if any, between outputs from this technique and those resulting from parallel multiple channel DOG processing of the input image. Unfortunately, owing to other priorities, this comparison has never been carried out. In the meantime, other forms of progressive multiple scale analysis have been evolved, based on statistical pooling of first or second differences of energy, motion, stereo disparity and various textural properties (see Chapters 9, 10, 15, 16).

11.6. Conclusions

It is possible, with a very small additional programming, to provide a noise-balanced facility for carrying out progressive serial scale reduction simply and cheaply using VISIVE-like processes. Such a progressive processing can use most of existing VISIVE functions and produce a series of data sets at various scales for vernier properties, profile feature analysis, motion and stereo disparity sensing. Alternatively, and equally simply, forms of progressive window analysis can be devised

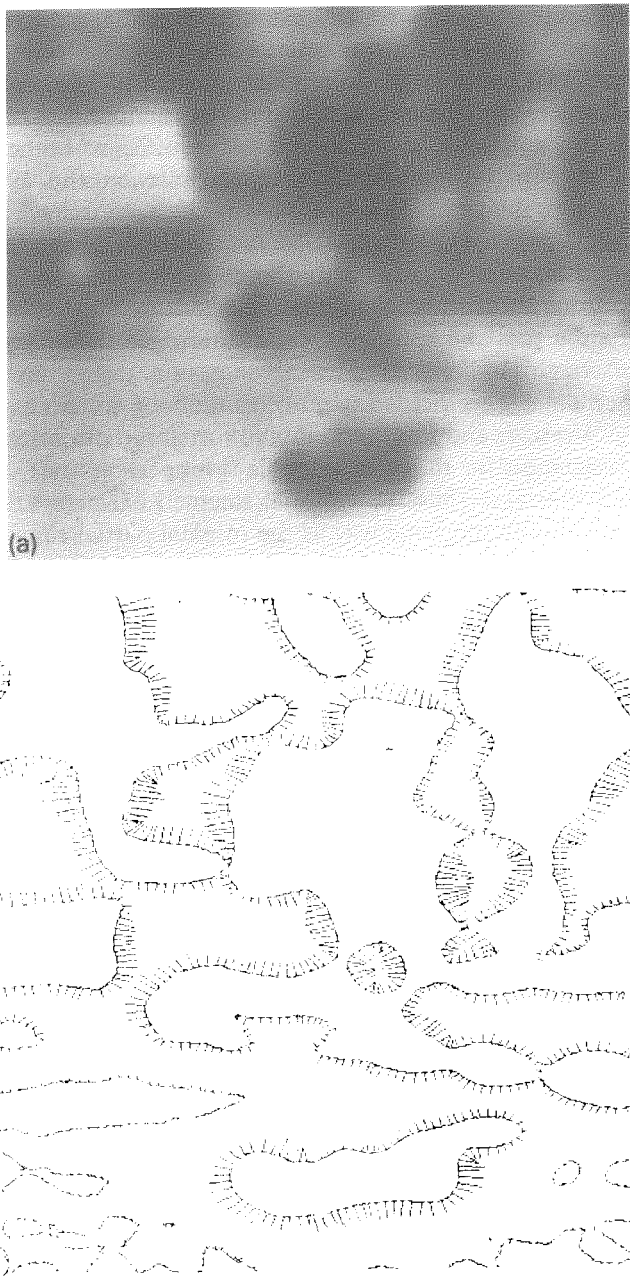

Fig. 11.4. A typical set of fragmentary profile maps for a set of VISIVE processes operating on the same scene fragment at progressive 2:1 scale reductions. In these figures the junction point of each 'T' represents the sub-pixel position, the cross-piece the local orientation and the length of the stem the relative strength of the profile fragment (see Chapter 4, Section 4.4 for a fuller explanation). [Note the evidence of the progressive loss of fine detail form on isolated objects and the coalescence of adjacent objects into single blobs]. a) Original scene. b) Profile map at original scale. c) Profile map at scale reduction of 2:1. d) Profile map at scale reduction of 4:1.

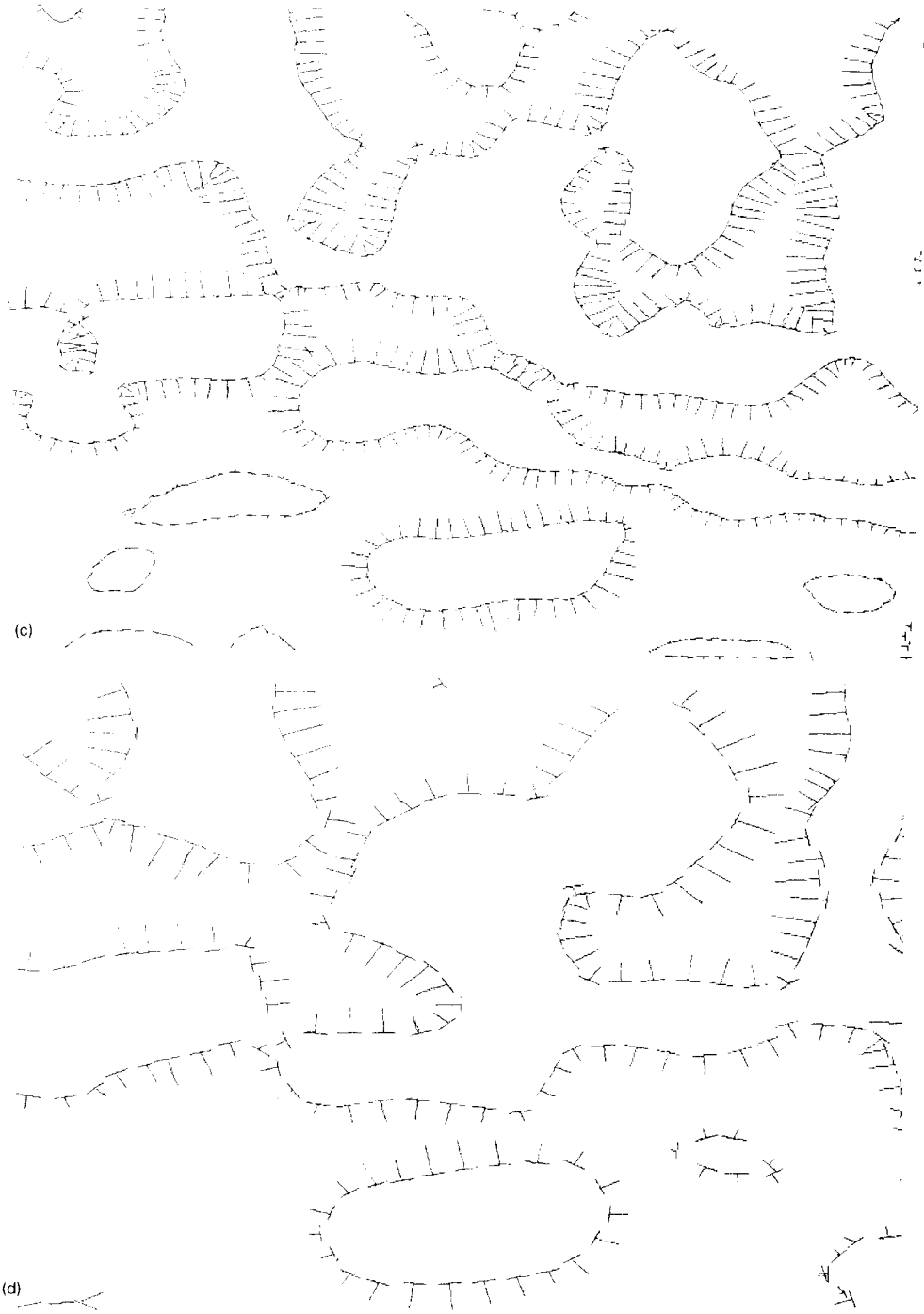

Fig. 11.4 (continued).

which operate by pooling any one of several attributes of local fragmentary profile data. It would appear that similar extensions could be applied to many other forms of single scale image processor. It may be that sensing of some scene properties (e.g. some forms of optical flow and stereo) are better carried out based on reduced scale analyses than by large scale statistics on the original, high fidelity outputs.

References

[11.1] Bracewell R.N. (1978), 'The Fourier Transform and its Applications (2nd Ed.)', McGraw Hill, New York.
[11.2] Overington I. (1976), 'Vision and Acquisition', Pentech Press, London. Chapter 2.
[11.3] Baker K. and Sullivan G.D. (1980), 'Multiple band-pass filters in image processing', IEE Proc., *127(Part E)*, 173.
[11.4] Overington I. (1982), 'Towards a complete model of photopic visual threshold performance', Opt. Eng., *21*, 002.
[11.5] Fiorentini A. and Maffei L. (1968), 'Perceptual correlates of inhibitory and facilitatory spatial interactions in the human visual system', Vision Research, *8*, 1195.
[11.6] Fiorentini A. and Maffei L. (1970), 'Transfer characteristics of excitation and inhibition in the human visual system', J. Neurophysiol., *33*, 285.
[11.7] Kulikowski J.J. and King-Smith P.E. (1973), 'Spatial arrangement of line, edge and grating detectors revealed by subthreshold summation', Vision Research, *13*, 1455.
[11.8] Westheimer G. (1967), 'Spatial interactions in human cone vision', J. Physiol., *190*, 139.
[11.9] Overington I. (1976), 'Vision and Acquisition', Pentech Press, London. Chapter 13.
[11.10] Bergen J.R. and Wilson H.R. (1979), 'A four channel model for threshold spatial vision', Vision Research, *19*, 19.
[11.11] Bergen J.R., Wilson H.R. and Cowan J.D. (1979), 'Further evidence for four mechanisms mediating vision at threshold: sensitivities to complex gratings and periodic stimuli', J. Opt. Soc. Am., *69*, 1580.
[11.12] Werblin F.S. (1973), 'The control of sensitivity in the retina', Scientific American, January, 71.
[11.13] Huck F.O. et al (1983), 'Information theory analysis of sensor-array imaging systems for computer vision', Proc. of the SPIE, Vol. 397, 82.
[11.14] Granlund G.H. (1983), 'Hierarchical image processing', Proc. of the SPIE, Vol. 397, 362.
[11.15] Granlund G.H. and Knutsson H. (1983), 'Contrast of structured and homogenous representations', in 'Physical and Biological Processing of Images' (Eds. O.J. Braddick and A.C. Sleigh), Springer-Verlag.
[11.16] Hunter G.M. and Steiglitz K. (1979), 'Linear transformation of pictures represented by quadtrees', Comp. Graph. and Image Proc., *10*, 289.
[11.17] Samet H. (1982), 'Neighbour finding techniques for images represented as quadtrees', Comp. Vis. and Imag. Proc., *18*, 37.
[11.18] Jackins C.L. and Tanimoto S.L. (1980), 'Octrees and their use in representing 3-dimensional objects', Comp. Grap. Imag. Proc., *14*, 249.
[11.19] Baroghimian G.A. and Klinger A. (1988), 'Space and time requirements for two image data structures', Proc. of the SPIE, Vol. 1002, 514.
[11.20] Ballard D.H. and Brown C.M. (1982), 'Computer Vision', Prentice Hall. Chapter 12.
[11.21] Kolb. H. (1970), 'Organisation of the outer plexiform layer of the primate retina: electron microscopy of Golgi-impregnated cells', Trans. R. Soc. Lond. B., *258*, 261.