

## ***IMAGE FEATURES***



# On scale and resolution in the analysis of local image structure

Kjell Brunnström, Jan-Olof Eklundh, Tony Lindeberg

Computer Vision and Associative Pattern Processing Laboratory (CVAP)\*

Royal Institute of Technology

S-100 44 Stockholm, Sweden

Email: kjellb@bion.kth.se, joe@bion.kth.se, tony@bion.kth.se

## Abstract

Focus-of-attention is extremely important in human visual perception. If computer vision systems are to perform tasks in a complex, dynamic world they will have to be able to control processing in a way that is analogous to visual attention in humans.

In this paper we will investigate problems in connection with foveation, that is examining selected regions of the world at high resolution. We will especially consider the problem of finding and classifying junctions from this aspect. We will show that foveation as simulated by controlled, active zooming in conjunction with scale-space techniques allows robust detection and classification of junctions.

## 1 Introduction

A central theme of computational vision is the derivation of local features of image brightness. Existing computational models often make reference to and are inspired by biological vision, see e.g. Marr and Hildreth (1980) or Watt (1989). From this perspective local feature extraction is similar to foveation. Usually, one also tacitly assumes that the goal of computer vision research is to develop methods for analyzing visual information that perform as well as the human visual system. In this paper we argue that such performance can hardly be expected by most current techniques, simply due to limits on resolution. Standard camera systems usually have a visual angle of about  $50^\circ$  and give image matrices of say  $500 \times 500$  up to  $1000 \times 1000$  pixels. This should be compared to the  $2^\circ$  in foveal vision, which in view of visual acuity can be said to correspond to an image resolution of about  $200 \times 200$  pixels. Obviously, this difference implies that multiresolution processing, like it is done in pyramids cannot be seen as analogous to foveation. The images coming out of a standard camera system are too much limited by resolution. In our experiments we overcome this limitation by doing controlled zooming and windowing. In this way we can simulate a system that performs foveation, that is *takes a closer look* at interesting regions in the field view.

An approach like this raises computational problems different from those appearing e.g. in edge detection or other general methods for searching for local structure in normal *overview*

---

\*The support from the National Swedish Board for Technical Development, STU, is gratefully acknowledged.

pictures. Obviously, we need a method for deciding where to focus our attention. This will be discussed in Sections 2–3. Moreover, the increased resolution is likely to enhance the noise at least relative to the prominence of the structures we are looking for. This may call for a different type of algorithms for detecting local structure. In fact, we will show how very simple local statistical techniques can be used to obtain robust detection and classification of junctions and corners or high-curvature points. Finally, we have the problem of detecting structure when we dynamically vary resolution and window size. What we propose is a method of stability of responses. The rationale for this approach is the observation that local structure, in the highly resolved foveated images, generally will belong to one of a small set of generic types. Stability of responses is therefore simple to assess. Notwithstanding this, there are more sophisticated approaches to the problems, as we will get back to later.

## 2 Detecting local structure

We will in this paper apply our idea of simulating foveation by active focusing to the detection and analysis of junctions of two or more boundary segments in gray-level images. We are particularly trying to find *T*-junctions and images of 3-dimensional corners, often showing up as junctions with three or more edge directions. These give important cues to 3-dimensional structure, for instance the *T*-junctions indicate interposition and hence relative depth. Naturally, also *L*-junctions or image corners are of interest.

It is well-known that elaborate edge detection methods like those proposed by Hückel (1971), Marr and Hildreth (1980), Canny (1986) or Bergholm (1987), have problems at junctions. Zero-crossings, that is boundaries between positive and negative regions (of the second derivatives) will, of course, not correctly divide three regions in a junction. If first order derivatives are used other problems arise, e.g. at the computation of the gradient direction.

To overcome such problems direct methods for junction detection have been proposed, e.g. by Moravec (1977), Nagel (1986) and Kitchen and Rosenfeld (1982). The results have been applied to matching for stereo and motion analysis, but it is not clear that these approaches give any precise information about the nature of the junction. On the other hand, explicit use of the feature as a cue to scene structure would require some sort of classification or description of the junction.

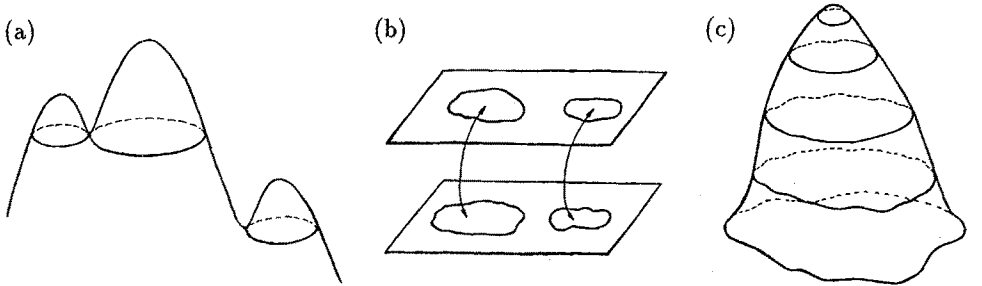
Proposed edge and junction detectors like those cited above are based on precise assumptions about the local image structure. Usually, one assumes two or three regions, each with locally constant or linearly varying intensities, regions which in turn are clearly separated. It is obvious that these models are difficult to assess from noisy data with very few samples from each population. Marr and Hildreth (1980) pointed out that while the smallest receptive fields according to e.g. Wilson (1983) contain about 500 cones, most edge and corner detectors work with a few tens of samples. This indicates that one should indeed work with large operators, as some authors also do, see e.g. Canny (1986) and Bergholm (1987). However, the problem is not solved with this. First, we still have the relation to the size of the structure we are looking for. We can't assess, say, a *three regions* model by simply using a large operator, if the window contains many different instances of it, just because the window is large. Hence, the idea of using large operators should only be expected to give *good* (see above) performance if it could be applied so that, at least ideally, no more than one feature occurred in each field. Now this poses a second problem. Since we only can assume coarse knowledge or expectance about the scale of the features how do we find a reasonable operator

size? Especially, how do we detect if two or a few nearby features appear in the same field? This can happen even if the resolution is increased. The proposed answer to this is to use the stability of the computed measures as the parameters, field size, resolution and location, are varied. More precisely, we have addressed the problem of finding the initial hypotheses about existing structure and its scale using scale-space representations. We will briefly review this theory in the next section. From these initial hypotheses zooming is performed and stability is verified or rejected.

### 3 Determining significant image structures and their scales

The prominent structures in the gray-level image can be determined using the scale-space primal sketch approach, developed in Lindeberg, Eklundh (1990) on the basis of the scale-space theory by Witkin (1983) and Koenderink (1984).

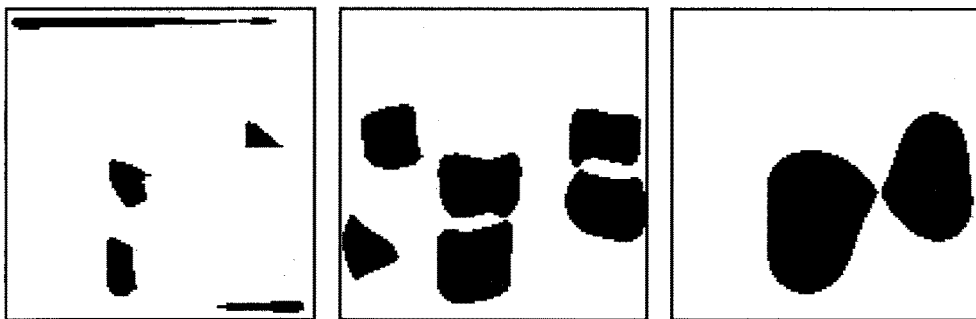
The main idea is to build a scale-space representation of blobs, i.e., peaks and valleys in the gray-level landscape and their *extent*. It is based on a conservative notion of a gray-level blob, which implies an inherent competition between parts, see Figure 1. These blobs are then linked between scales into higher order objects, called scale-space blobs, which have extent not only in space and gray-level but also in scale. The bifurcations occurring in scale-space are registered and a hierarchical data structure is created. There are several computational problems to solve in building this representation, we refer the reader to Lindeberg, Eklundh (1990) for the details.



**Figure 1:** (a) Example illustrating the definition of a gray-level blob. (b) By linking gray-level blobs between scales one obtains (c) scale-space blobs which are objects having extent both in space, gray-level and scale. These objects are the fundamental primitives in the representation.

From measurements of significance and stability in scale-space the representation gives a qualitative description of the image structure, which enables determination of approximate location and extent of relevant regions in the images as well as appropriate levels of scale for treating those regions. The significance value of a scale-space blob is essentially the volume of the scale-space blob in the four-dimensional scale-space given by the spatial, gray-level and scale coordinates. See Figure 2 for an example.

Experimentally it has been demonstrated that the extracted information correspond to perceptually relevant regions in the image. Hence, we suggest the use of this output for generating hypothesis for the focusing procedure as well as for controlling the setting of initial values for window size and position.



**Figure 2:** Illustration of the scale-space primal sketch output for the upper right block image in Figure 3. The dark regions correspond to significant blobs in the scale-space primal sketch. Note that they are not just plain regions in the image, but entities with well-defined scale information and an associated significance measure determined from the behaviour of the blob in scale-space. Observe that the individual blocks are extracted as single units and that the adjacent blocks become grouped into coarser scale objects.

It should be stressed that the intrinsic shape of the gray-level landscape is extracted in a completely bottom-up data-driven way without any a priori information about spatial location, scale or shape of the primitives.

An important problem in deriving this representation concerns measuring significance and scale in such a way that comparisons between significance values can be made for blobs existing at different levels of scale. This issue must be explicitly dealt with. For instance, a concept called "effective scale", which is the natural unit for measurements of scale-space lifetime, needs to be introduced. It is argued, that when the effective scale is increased with a small increment then the expected amount of image structure to be destroyed as well as the probability that an extremum point disappears should be independent of the current scale and the current amount of structure. From these requirements we have shown that this effective scale can be defined in, in principle, one way only. The resulting effective scale is essentially the logarithm of the expected number of local extrema at the given level of scale, see Lindeberg, Eklundh (1990) for the details as well as the difficult question about how to estimate the expected amount of structure in an image.

The significance values obtained from scale-space primal sketch also induce a well-defined ranking between the blobs, which empirically has been found to give an intuitively reasonable ordering of the events. Thus, we believe that such a module really can serve as a guide to the focus-of-attention.

## 4 Image structure at junctions

We will now turn to our problem of finding and classifying junctions. In the next section we will briefly review a technique for their detection. It turns out that most methods working on standard imagery will have thresholding problems. However, various additions to the basic algorithm can be used to limit the number of falsely detected corners, with a certain risk of loosing some of the true feature points. However, this trade-off seems inevitable and tends

to occur in all proposed approaches. In particular, false responses tend to occur along edges. Also noise spikes, if they exist, will according to the model give responses that are strong local maxima.

There are, in fact, five situations in which strong responses mainly occur:

at noise spikes
along edges
at 2-junctions
at 3-junctions
at $n$ -junctions, $n > 3$

Let us now consider an ideal case, in which the resolution is infinite and each region corresponds to a smoothly varying surface. Let us also assume that the illumination varies smoothly over each region. In that situation we could easily discriminate between the different cases by considering the distributions of gray-levels and edge directions in a sufficiently small neighbourhood of the possible (image) corner. In fact, this would be possible also if the surfaces contained some non-infinitesimal surface markings as well. The classification would be:

Case	Intensity	Edge direction
at noise spikes	uniform	*
along edges	bimodal	unimodal
at 2-junctions	bimodal	bimodal
at T-junctions	trimodal	bimodal
at 3-junctions	trimodal	trimodal
at $n$ -junctions	*	*

where — \* stands for inconclusive.

Experiments by Nagel and his co-workers, Nagel (1989), show that high quality intensity images indeed satisfy such conditions just as a straightforward model would predict.

The important question now is what happens in a realistic case, e.g. with direct and indirect illumination and noise and with finite resolution. One would still expect, as is indeed the basis for most of the low-level methods propose, that the given classification would be valid anyway. However, establishing this classification requires sufficiently many samples of the different distribution. Moreover, there is a need for a method of classification that is robust with respect to noise and variations due to the imaging process.

What this means in practice is that the resolution has to be high enough and that the window size should be appropriately chosen for correct classification. If we had a precise model for the imaging process and the noise characteristics, one could conceive deriving bounds on the resolution and the window size, at least in some simple cases. However, apart from the fact that this might be difficult in itself, such models are hardly available. What we propose instead is to use the process of focusing. Focusing means that we increase the resolution locally in a *continuous* manner (even though we still have to sample at discrete resolutions). We can simultaneously either keep the window size constant with respect to the original image, or vary it independently. Generally we want to decrease the window sizes. The contention we make about the focusing approach is that it avoids the problem of selecting an appropriate pair of resolution/window size, since the simple features we are looking for will show up in a stable manner during the variation of the parameters. An important reason behind this argument is that the classification can be based on simple features (the number of peaks of

a histogram) of simple statistics. We can summarize the main principles of the approach as follows:

- *take a closer look* at candidate interest points by increasing the resolution, i.e. acquiring a *new image* with higher sampling density, and varying the window size
- detect stable classifications based on the simple features in the table above

So far, we have used no objective method to assess what we mean by *detecting the stable classifications*. However, we will show a number of experiments that demonstrate that our computational model indeed predicts what happens. Moreover, the approach described in Section 3 can be applied also to this problem.

## 5 Experimental technique

We will demonstrate our approach on a set of images at different resolutions of scenes containing simple man-made objects. Let us describe the processing steps.

We first perform a bottom-up pre-processing step, in which a set of candidate points are determined, as described in Sections 2 and 3. These points are tentative junctions, that is points where several regions meet. A crucial issue is to find a set of such points that is limited in size and is likely to contain true and significant junctions in the image.

The algorithm we use is based on the method of computing directional variances, suggested by Moravec (1977) and contains some additional constraints of *cornerity*. In this way significant junctions can be automatically obtained without any strong dependence on thresholding, with is difficult in Moravec's original algorithm. We will not discuss this further since we are not trying to analyze all junctions here. See Brunnström et al (1989) for details.

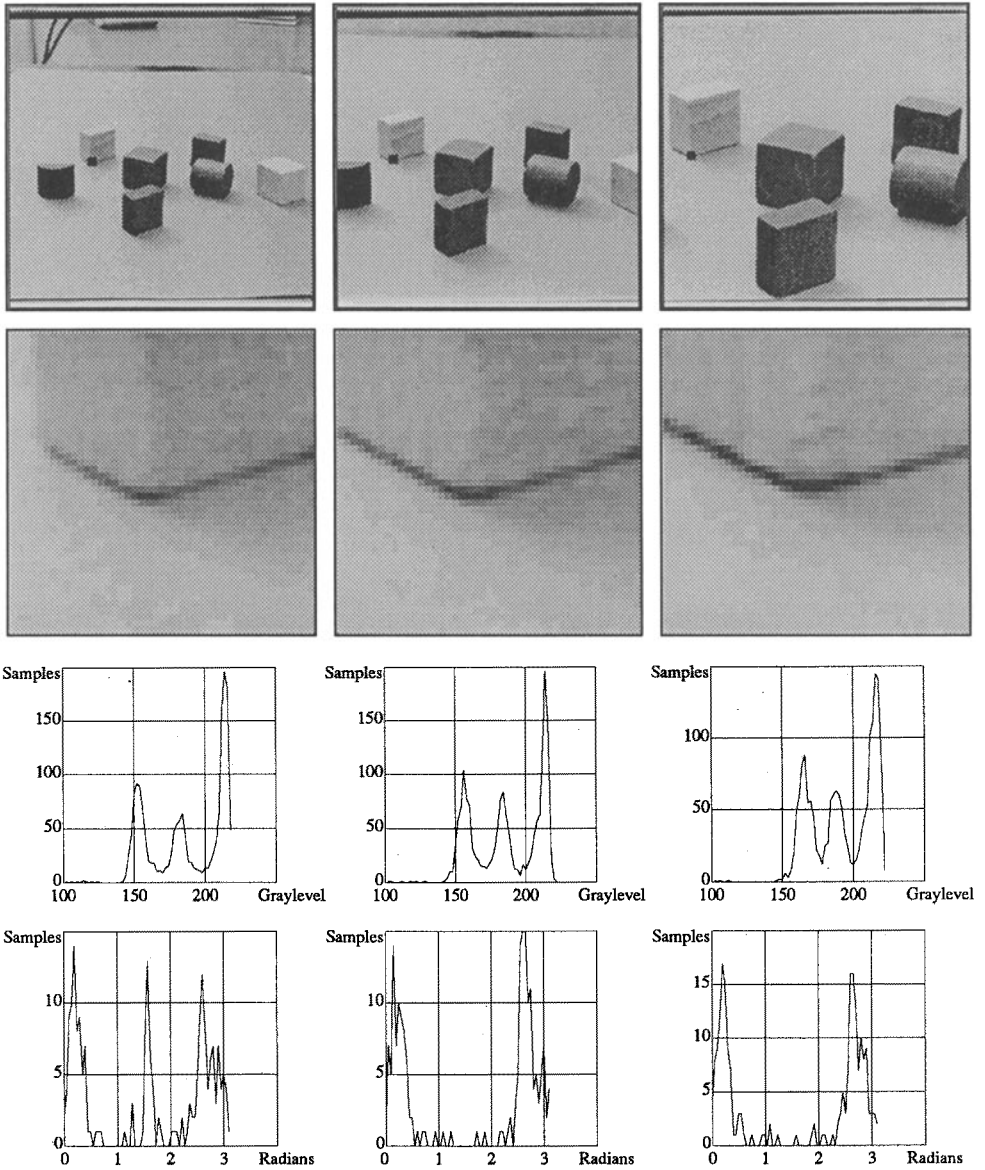
Having found a point of interest, we next simply complete the distribution of gray-levels and edge directions in a sequence of windows of varying resolution and size around the point. The edge direction statistics are obtained for pixels labelled as edge elements only. Hence, some edge detector has to be applied first. A simple detector like Roberts' cross might do, but the choice is of little importance. Something that may be important, though, is the effect of point of interest itself. As discussed above it is at the point that most low-level algorithms will have problems. Assuming that the resolution is high enough, so that we can appropriately sample the postulated simple underlying structure in the neighbourhood, we can treat the point as being a removable discontinuity. Hence, we compute the statistics in all but a subwindow around the point. Since we vary both the resolution and the window size and look for stability, we have chosen to let the subwindow cover a fixed part of the total window, say  $\frac{1}{9}$  or  $\frac{1}{16}$  of the area.

The decisions to make only concern how many peaks there are in the histograms, and if these numbers are stable and fit with any of the generic cases. We have used a method for sharpening the peaks in histograms proposed by Peleg (1978). This method is basically equivalent to smoothing and peak detection.

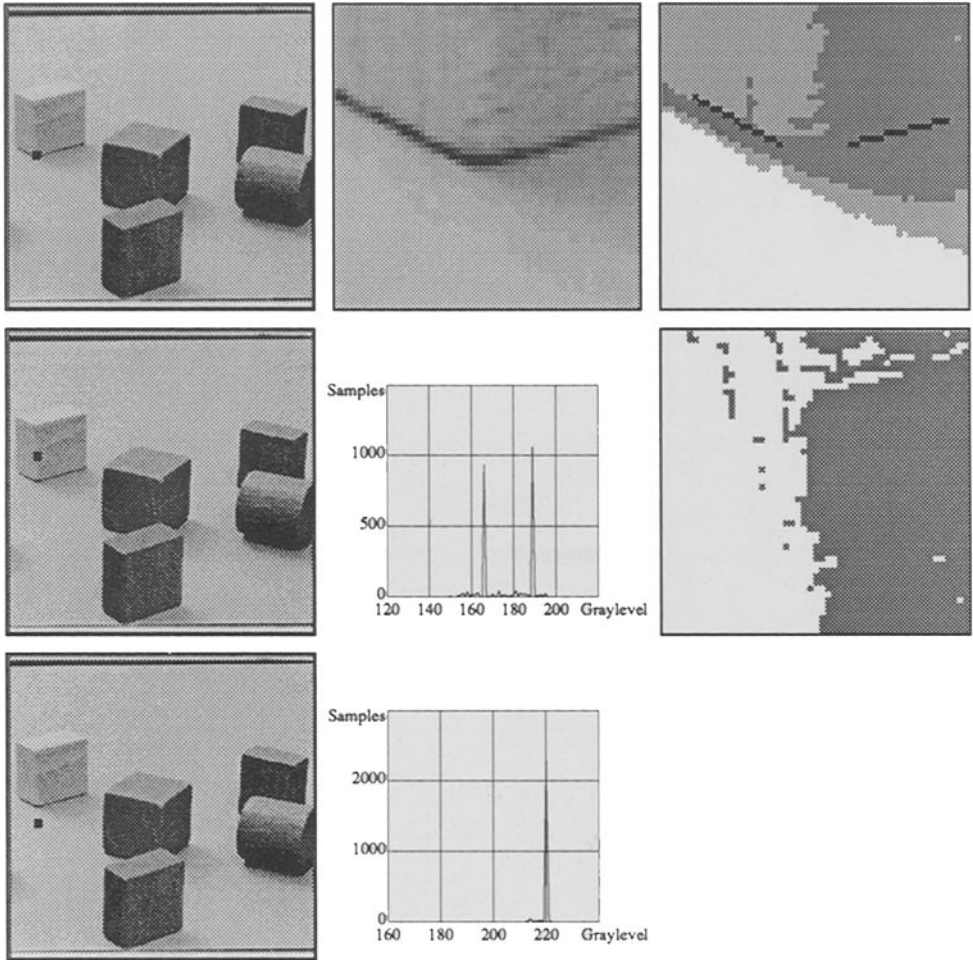
## 6 Results

In the toy block image in Figure 3, we focus on a corner of the bright block to the left. The possible existence of a corner has previously been established according to Section 5.





**Figure 3:** The figure illustrates the stability of the local statistics under resolution variations. In the left column, where the coarsest resolution is shown, the window is too large, which here shows up as an extra peak in the edge direction histogram.

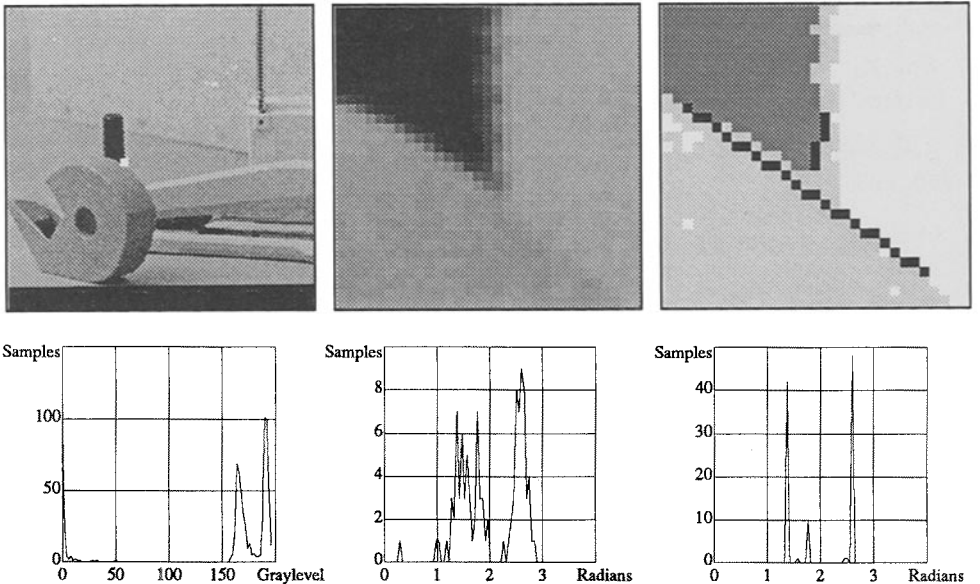


**Figure 4:** The analysis of the fixation point shown in the top left image, with a neighbourhood size shown in the top middle window, gives the result shown at the top right, i.e., three gray-levels and two edge directions (indicating an L-junction). This is not a generic case. For resolving the situation two new fixation points are analyzed, shown in the middle and in the bottom row. This results in the correct classification.

At coarse resolution three gray-levels and *three* directions are detected. This is because the outer boundary on the left is visible. As the focusing procedure continues, there is a stable response of three gray-levels and *two* directions. The vertical edge is very weak and cannot be detected, not even by e.g. Canny–Deriche’s edge detector. However, backprojection into the image shows that this response is not generic. Therefore, it is assumed that some information is missing or incorrect. The point focused at is therefore shifted up and down the bisector of the detected L-junction, in a simulated eye-movement. The correct structure can then be assessed, as is illustrated in Figure 4. Note that in all cases both the window size and the resolution is varied, but only very few pictures can be shown here.

It should be stressed that the results shown in Figures 3–4 are obtained with simple tests in a case when e.g. sophisticated edge detectors fail.

Figure 5 shows some illustrative output in the case of a true T-junction and when one of the objects is curved. We are showing the final stable results which indicate a T-junction. Notably, a third but weakly supported direction is detected. This direction is, in fact, due to the lower part of the curved edge. However, backprojection of the predicted edges can again be used to characterize this type of response.



**Figure 5:** A T-junction between a slightly curved contour and a straight one. The left image shows the overview of the scene, the middle shows the local neighbourhood which has been processed and the right shows the result. The histograms are from left to right; gray-level, edge direction and peak sharpened edge direction.

## 7 Conclusions

We have argued that focus-of-attention mechanisms are necessary in computer vision systems, if they are to perform their tasks in a complex, dynamic world. We discussed how foveation,

that is examining selected regions of the visual world at high resolution, can be incorporated in an active vision system. We studied the use of this capability in the task of finding and classifying junctions of edges. It was shown that the junctions in this context could be labelled into a small set of generic cases and that these cases could be established computationally by robust and simple algorithms.

## References

- [1] Bergholm, F., (1987), *Edge Focusing*, IEEE PAMI, 9:6, 726-741
- [2] Brunnström, K., Eklundh, J.O., Kakimoto, A., (1989), *On Focus-of-Attention by Active Focusing*, Proc. NATO ASI on Robotics and Active Computer Vision, Springer Verlag, New York, in press.
- [3] Canny, J.F., (1986), *A Computational Approach to Edge Detection*, IEEE PAMI, 8:6, 679-698
- [4] Hückel, M., (1971), *Operator which Locates Edges in Digitized Pictures*, JACM, 18, 113-125
- [5] Kitchen, L., Rosenfeld, R., (1982), *Gray-Level Corner Detection*, Pattern Recognition Letters, 1:2, 95-102
- [6] Koenderink J.J., van Doorn A.J. (1984) *The Structure of Images*, Biological Cybernetics, 50, pp363-370.
- [7] Lindeberg, T., Eklundh, J.O., (1990), *On the Computation of a Scale-Space Primal Sketch*, submitted
- [8] Marr, D., Hildreth, E., (1980), *Theory of Edge Detection*, Proc. Royal Society of London, B-207, 187-217
- [9] Moravec, H.P., (1977), *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*, Stanford AIM-340
- [10] Nagel, H.H., (1986), *Image Sequences — The (Octal) Years — From Phenomenology towards a Theoretical Foundation*, Proc. 8th ICPR, 1174-1185
- [11] Nagel, H.H., (1989), *Personal communication*
- [12] Peleg, S., (1978), *Iterative Histogram Modification 2*, IEEE SMC, 8, 555-556
- [13] Watt, R., (1988), *Visual Processing: Computational Psychophysical and Cognitive Research*, Laurence Erlbaum Associates, Publishers, London
- [14] Wilson, H.R., (1983), *Psychophysical Evidence for Spatial Channels*, Braddick, O.J., Sleigh, A.C., eds., in *Physical and Biological Proc. of Images*, Springer Verlag, New York
- [15] Witkin, A.P., (1983), *Scale-Space Filtering*, Proc. 8th IJCAI, 1019-1021