This is the published version of a paper published in *International Journal of Computer Vision*.

Citation for the original published paper (version of record):

Active fixation for scene exploration
*International Journal of Computer Vision*, 17(2): 137-162

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-24525
Active Fixation for Scene Exploration

KJELL BRUNNSTRÖM, JAN-OLOF EKLUNDH AND TOMAS UHLIN
Computational Vision and Active Perception Laboratory (CVAP), Department of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology), S-100 44 Stockholm, Sweden
kjellb@bion.kth.se
joe@bion.kth.se
tomas@bion.kth.se

Received March 1993; Revised February 1994 and September 1994

Abstract. It is well-known that active selection of fixation points in humans is highly context and task dependent. It is therefore likely that successful computational processes for fixation in active vision should be so too. We are considering active fixation in the context of recognition of man-made objects characterized by their shapes. In this situation the qualitative shape and type of observed junctions play an important role. The fixations are driven by a grouping strategy, which forms sets of connected junctions separated from the surrounding at depth discontinuities. We have furthermore developed a methodology for rapid active detection and classification of junctions by selection of fixation points. The approach is based on direct computations from image data and allows integration of stereo and accommodation cues with luminance information. This work form a part of an effort to perform active recognition of generic objects, in the spirit of Malik and Biederman, but on real imagery rather than on line-drawings.

1. Introduction

Biological vision is inherently active. By visual perception humans and animals actively acquire information from the environment to guide their behavior. In recent years active computer vision has attracted interest as a paradigm for studying and developing “seeing systems” (Bajcsy, 1985; Aloimonos et al., 1987; Pahlavan et al., 1992). The emphasis of the work has been on real-time gaze control and although the crucial importance of fixation has been stressed e.g. by Ballard (1991), only limited work has been done to investigate computational strategies for choosing fixation points. Instead most efforts have been devoted to reactive gaze control. Exceptions exist, like in the work of Bajcsy and Campos (1992), which discuss a general framework of control in what is called the “where to look next” problem and in Rimey and Brown (1992), who apply a probabilistic framework to address the same problem.

In biological vision the fixation patterns are task dependent. It is likely that useful strategies for actively choosing fixations should be so also in robot vision. The cited works involve such a dependence. They put emphasis on inference mechanisms that can lead to a solution of the specific task, which in both cases is of high level nature. In this paper we consider a task more closely related to the visual processes. Our goal is to demonstrate the role of choosing fixations and attentional points in the actual derivation of useful information from the scene. To be able to do so without a comprehensive model of the seeing agent we define a scenario and a task of limited complexity.

The underlying task is that of recognition of manufactured objects characterized by their shape. In this case important cues are provided e.g. by junctions, and their types, and edges, in particular classified as being straight and curved. Our goal in this work is to actively obtain such information by an attentional step followed by a set of fixations. The approach has some relations
to the studies made in Culhane and Tsotsos (1992), Westclius et al. (1991). However, our work aims further than just determining fixation points insofar that it attempts to provide a rapid categorization of features as well. Moreover, it relies on fixation in the scene and can therefore use cues to depth like binocular disparity and focus. In particular, this allows us to include a process corresponding to foveation, that is acquiring information at very high resolution around a point of interest in the world. Finally, the approach does not require that all the interesting parts of the scene are in the field of view initially.

Recent work on how a mobile observer can choose informative viewing directions can be found in Wilkes and Tsotsos (1992, 1993b). Although this work has a similar goal in object recognition, the work is based on standard static methods for obtaining the features on which the recognition is based, and instead focuses on from where to look at an object rather than on what. Moreover, Dickinson and colleagues (1992a, 1992b) have considered aspect oriented approaches in model-based static object recognition. Again, this work does not investigate what the actual features are, but instead quickly tries hypotheses of what is seen, relying on features that are statically obtained with standard methods.

1.1. Problem Domain

The study of active vision systems in general is a formidable task, since it requires modeling the seeing agent. We have therefore chosen to select a very specific problem domain, which is rich enough to serve as a test of our hypotheses, and which is also of importance in applications of computer vision. The chosen domain is a static world containing man-made objects. This domain has been appearing in many studies i.e. in line-drawing analysis, the difference here being that this world is regarded by an active observer.

Although simple, the chosen scenario is important due to its potential applications in robot vision, and also, as proposed by Biederman (1985, 1987), because it has strong implications on recognition in general. At the same time this world is so simple that we can model the goals and intentions of the seeing agent as being that of exploring the world for all occurrences of the object it knows. The strategies will then be determined from what it knows about these objects, in this case the edge-junction structure, which we can consider as being its long term memory.

It is well-known that a major problem in this context is to find and characterize appropriate local features for performing the analysis at the object level. This problem comprises both the ubiquitous figure-ground segmentation problem and perceptual grouping. Here we want to address these issues from an active vision perspective, using selective fixation.

To be more precise, our active observer has binocular vision and is capable of fixating points in its three-dimensional environment. Its visual strategies are, as mentioned above, entirely determined by its knowledge about the objects it is looking for. What we want to show then, is that

- The active observer can generate groups of local features that are meaningful in the scene, in the sense that they suggest possible objects. In particular this means that although fixation is changed i.e. that different partial views of the scene are acquired, a coherent view of the objects in the world is obtained.
- This process does not require interpretation or even processing of all information available from the scene.
- The process provides sparse but generally appropriate groups of features. Hence it can avoid potential combinatorial explosion by using attention.
- The active observer requires and benefits from the integration of many cues, monocular and binocular (motion is not used here but could of course be included).
- The active observer relies on data from a retinotopic front-end of a simple and parallel nature. This suggests that it can be given "fast" implementation, but also puts restrictions on the operations allowed.

These hypotheses are illustrated and tested in a general framework for selective fixation based on an attentional step followed by a set of foveations2. In the actual implementations various techniques have been developed and employed, which however should not be seen as the major contribution of this work. This is instead the overall framework and the extent to which it supports our underlying general hypotheses.

A final hypothesis, that in a sense permeates and forms a basis of several of the previous ones is that the active observer can and should rely on a combination of foveal and peripheral vision.

A word on the role of eye movements in biological vision is appropriate here. They are usually classified as follows:
Eye movements that assist a stable perception.
Eye movements due to the foveal retinal structure.

In the case of active machine perception and sensors, with uniformly distributed resolution, it is obvious that our camera movements accompanied by zooming correspond to saccadic eye-movements and in this way, these movements give us access to a wider field of view and a higher resolution. The lower resolution in the region that is not zoomed in can also be considered as the correspondence to perifoveal or peripheral retinal areas, helping an easy detection of potential regions of interest.

To appreciate foveation, it is essential to notice the great difference between our most advanced imaging sensors and the human retina. Many of the traditional difficulties in the low-level image processing would be markedly simplified just due to a higher resolution; a higher resolution that decreases according to a scheme that is temporally and spatially fitted to the significance of the patterns and events of interest.

The standard argument against the role of resolution in a successful classification is often the fact that human beings can apparently detect different structures in low resolution images without a great effort. The counter argument, that this is done easily only in the case of known objects and structures, is however valid as well. Consequently, addressing the problem of junction detection and classification from a pure low-level point of view is still justified and significant.

1.2. Fixation vs Attention

The distinction between the words fixation and attention needs to be pointed out. Many authors use the word fixation, when attention is what is really meant. Fixation means a (mechanical) movement of the eyes (cameras), where at the end both eyes are pointed to and accommodated at the same point in space. Attention on the other hand is the current region of interest for the processing, which means that as long as the proper information is available, several areas can be attended to at a single fixation. If a very small region in the center is processed at each fixation these two terms become synonymous.

1.3. Objects and Features

Classical work in line drawing analysis from Waltz (1975) to Malik (1987) has stressed the importance of junctions for constraining the possible interpretations. Malik also provided a catalogue of possible appearances of junctions in a perfect line drawing, see figure 1. Biederman (1985, 1987), has shown that humans can do rapid model indexing on the basis of qualitative cues of a similar nature, including edges classified as straight or curved.

Work on computing such features from realistic images is abundant. The most common approach is to do edge detection followed by some linking procedure. Various methods are then used for classification, e.g. approximation techniques. In our view such approaches do not fit well with the notion of an active vision system. The tracing and linking step is based on local decisions while the classification which requires global or at least multilocial coordination, is done by indirect bottom-up computations. What we want to study is an active approach including an attentional mechanism and selective fixation. By such a technique we arrive at a visual routine, in the sense of Ullman (1987), which rapidly can pick up sufficient information to detect, localize and characterize the features we are looking for, in this case junctions and their edge types. The only information used is what is obtained from a visual front-end in the sense of Koenderink and van Doorn (1990), or from sets of directionally sensitive local filters. Hence the computations can be seen as direct. In active vision, depth cues from binocular disparities or accommodation are available as well. It will be shown how accommodation cues can be integrated in the classification scheme and that this could be done with binocular disparities as well.
As mentioned above the objects in our problem domain are man-made objects characterized by their shape. This is a rather vague definition. Malik (1987) in his study of line drawings defined objects by

...opaque solid objects bounded by piecewise smooth surfaces with no markings or texture.

We will use this as our definition of objects. In reality objects totally without texture and cracks, do not exist, but as long as these effects are moderate they could be handled using algorithms with robust properties. On the other hand this world is richer than the polyhedral world which it includes. Moreover, objects modelled by geons will also be covered with this definition. Such objects in a perfect line drawing can be interpreted using a small set of junctions. Figure 1 shows their basic appearance as they were suggested by Malik. Furthermore, Malik has shown that, close to a junction point they can be approximated to a first order with $L$, $Y$, $T$- and $T$-junctions.

The outline of the remaining paper is as follows. We first briefly describe our earlier results on which this work is based. We then describe the computational principles of our approach, including the integration of additional cues. Finally we describe our experimental results and end with a discussion of our intended integrated system.

2. Earlier Work

In Brunnström et al. (1990, 1992), we have described how we can find junctions and classify them with regard to how many edges and regions that meet there. This is done by combining an attentional step, finding regions of interest, with a step simulating foveation. Two important points were stressed in this work. One was that the "foveation" step, simulated by computer controlled zoom, by providing high resolution data helped to overcome many problems typically occurring in feature detection. The other point was that a multi-scale representation of the wide-field original images could be used to guide the attention. In fact, the approach used, based on multi-scale computations of high curvature points and developed by Lindeberg (1993, 1994), also allows determination of the scales at which the features best could be defined and described.

In this work we also presented an integrated active system for performing such a chain of operations, although without any particular connection to a high-level visual task and hence without any strategy for choosing next fixation or combining information at different locations. What we will describe in this paper is a technique for selecting and grouping fixation points in the scene using information extracted at already attended points and classify the features into categories useful in object recognition of the type mentioned in the introduction.

3. Computational Aspects

As already mentioned our goal is to actively perform grouping and classification of junctions. This is achieved by a sequence of steps which involve:

- Initially to find some structure to attend to.
- The ability to fixate.
- The strategy to choose the next attention point and decide whether a re-fixation is needed.
- The decision whether two junctions should be connected.
- The ability to relate structure from one fixation to another.

We will in this section describe these steps in more detail.

Before getting into details we will outline the grouping process briefly. The clustering of junctions is started with a classified junction. The strategy is then to re-fixate at the end of one leg, and search for a new junction compatible with the already found edge-curve. A new classification is performed at the new fixation. The legs corresponding to the same curve are connected. The search is continued at the end of one of the unconnected legs. The other legs are queued until later. When a $T$-junction is established, the leg which does not match the leg of the last fixation will be discarded. If it is the start junction of an object, the search of the occluding edge will be given priority if no other information is present. The junctions will be connected through their matching legs.

The exploration can be initiated by starting at a point indicated by an attentional process for instance as discussed in our previous work or simply by picking a highly significant junction candidate, obtained with an interest operator, e.g. (Moravec, 1977; Kitchen and Rosenfeld, 1982; Dreschler and Nagel, 1982; Förstner and Gülich, 1987), which is then classified. The classification method, which will be described later in this section, establishes the number of curves meeting at
the junction, their type—straight or curved—and also tries to find the extent of the curves. This is done using an active approach, which means that zoom, focus and stereo are incorporated in the analysis as well.

In order to distinguish straight or curved curves meeting at the junction a semi-global part of the edge must be taken into account. This is obtained by moving the attention point in the expected direction of the edge for a small number of steps and at each attended region locate the edge.

The search along the edge is based on the "predict-match-update" strategy. Taking a predicted direction the attention point will be moved along it. A search for the curve position is performed and the local tangential direction measured. Two different strategies will be discussed here—one simple and straightforward and one based on a Kalman filter. The first method is given here because it illustrates the principle very well. However, to increase robustness it is necessary to measure several characteristics about the curve at each attention. The Kalman filter gives a common framework to handle this.

Initially at the junction an estimation of dominant directions in its immediate surroundings has to be performed for starting the process. The method used here will be described later in this section.

### 3.1. Changing Attention Point Along a Curve: Straightforward Approach

In this section we will describe a straightforward way of implementing a process, which moves its attention point along a curve. Starting with a position and a direction a new predicted direction is calculated by adding the difference in angle from the last measurement to the just measured one. The image will be explored in the predicted direction, by moving the attention point a step in that direction. The step length is a variable which starts at a low value and will be increased at each step, unless the process fails to find the predicted direction at the new location, at which the step length is decreased, as illustrated by figure 2. The measurement position of the new direction will be searched for in a sector formed by an uncertainty angle and the predicted direction. The point which has the maximum gradient magnitude and a direction within tolerance will be taken as the new measurement.

If we call the $n$:th measured position of the curve $p_n = (x_n, y_n)^T$ and the direction is represented with the unit vector $v_n$, then a new predicted direction and position of the curve can be computed by

$$
\hat{v}_{n+1} = \hat{v}_n + C \cdot (\hat{v}_n - \hat{v}_{n-1}) \quad (1)
$$

$$
\hat{p}_{n+1} = \hat{p}_n + l_n \cdot \hat{v}_{n+1} \quad (2)
$$

where $C$ is a gain constant, which in these experiments has mostly been kept equal to one. $l_n$ denotes the current step length, which is updated according to

$$
l_{n+1} = l_n \pm \beta \cdot s \quad (3)
$$

where "+" is used if the curve is found and "-" otherwise. However, it will never be less than $l_0$; the start value of the step length. $\beta$ and $l_0$ are preset constants and $s$ is a measure of a characteristic length in the neighborhood of the junction, that is a scale measure at object level.

The search area, $S$, at an attention point with the position $\hat{p}_{n+1}$ is given by

$$
S = \{ q : |\arg(q - q_0) - \arg(\hat{p}_{n+1})| < \Delta \theta \text{ and } |q - \hat{p}_{n+1}| < R_0 \} \quad (4)
$$

$$
q_0 = \hat{p}_{n+1} - R_0 \cdot \hat{v}_{n+1}
$$

![Figure 2. Starting with a direction prediction (1) a new attention point is calculated (2) in the direction of the prediction with the current step length. The attention area is illustrated with a circle. The area between the two lines and the circle, see the enlargement, will be searched for a point with compatible direction and high gradient magnitude, see the enlargement. If no such point is found (3) the step length is decreased and a new search is performed (4). The process tries to increase the step length for each new successful attention. The enlarged circle (with radius $R_0$), centered at the predicted location ($\hat{p}$), shows how a search area is formed between two lines, symmetrically placed around the predicted direction ($\hat{v}$). The opening angle between the lines is $\Delta \theta$ and they intersect each other on the circle.](image)
where \( R_0 \) is the radius of a circle centered at the attention point and \( q \) is any point inside it. \( \Delta \theta \) is the half angle between the lines restricting the search area, see figure 2. In \( S \) the curve position, \( \hat{p}_{n+1} \), is given by that point \( q \), which fulfills the following conditions

\[
\max \| \nabla I(q) \|, \; q \in S \quad \text{and} \quad |\arg(\nabla I(q)) - \arg(\hat{p}_{n+1})| < \Delta \theta
\]

where \( \nabla I(\cdot) \) stands for the gradient of the image function \( I(\cdot) \). That is, we maximize the gradient magnitude subject to the condition that the gradient direction is close to the predicted direction. We can now calculate our new direction estimate as

\[
\hat{v}_{n+1} = \frac{\nabla I(p_{n+1})}{\| \nabla I(p_{n+1}) \|}
\]

There are two vectors that satisfy Eq. (6). The direction closest to \( v_n \) is chosen.

3.2. Changing Attention Point Along a Curve: Kalman Filtering Approach

The problem of moving the attention point along a curve can be solved with a “predict-match-update” strategy. A powerful tool to handle this type of process is a Kalman filter, divided into a measurement and a prediction phase. Apart from the optimality of the filter under certain assumptions, it gives a unified treatment of different types of measured entities in the process, such as location, direction, contrast, scale, depth etc. The process is roughly described in figure 2. Given an initial estimation of the curve, a prediction of its location and direction are obtained from the filter. These predictions are used to define a search area in which the location of the curve is searched for. The measurements made at this location are used in the filter update. This predict-update procedure is repeated as long as consistent measurements of the curve are found. A detailed description of the Kalman filter is presented in Appendix A. The search area is defined in the same way as for the straightforward method, see Eq. (5) and figure 2.

The tangent of the curve could be approximated with the direction at a local maximum of the gradient magnitude as described in the last section, but in order to increase robustness and make a better estimation of the tangent of the underlying curve, a line fitting procedure will be employed. We pick out some local maximum points of the prediction weighted gradient magnitude, which is defined by

\[
P(q) = \| \nabla I(q) \| \cdot \cos^n(\arg(\nabla I(q)) - \arg(\hat{v}_n)),
\]

where \( \hat{v} \) stands for the predicted direction. This is done by searching for max points perpendicular to the predicted direction at different distances from \( q_0 \). A line is fitted to these points with a RANSAC algorithm (Fischler and Bolles, 1981), with the extra condition that \( P(q) \) should be close to constant for these points. The direction is then given by the direction of this line and the contrast will be represented by the mean of \( P(q) \) of these points. The width of the edge can be found by computing the covariance matrix for the spatial distribution of values of \( P(q) \) weighted with its power in this neighborhood, as described in Section 3.4. The measured quantities are then fed back into the filter. For an example see figure 4.

3.3. Finding Initial Direction Estimates

First the dominant directions in the neighborhood are found by forming a directional histogram and the dominant peaks are obtained by a peak sharpening procedure. The histogram has been formed in such a way that the peak position gives the dominant directions directly, which could be used as start values. However, as with all voting techniques there is a trade-off between having large accumulators for good detection and keeping them small for precision in the measurement. A well known way to overcome this problem is to add a localization steps after the detection. These two step will be described in some detail below.

The edge direction histograms can be formed by just considering the distribution of edge directions and this will in most cases be sufficient. However, this does not use the fact that a junction has a very special geometry. Ideally each edge pixel will have a gradient direction which is normal to the direction from the junction point to the pixel, here called the positional directions. So instead of accumulating the gradient directions, the positional direction modulo \( 2\pi \) are accumulated weighted with the inner product of the positional direction and the direction orthogonal to the gradient. If we denote the vector from the junction to a pixel in the neighborhood
with \( p \), then each vote is weighted with

\[
W(p) = w(p) \left( e^{-\frac{p^T \nabla I(p) \cdot p}{\sigma_1^2}} - e^{-\frac{p^T \nabla I(p) \cdot p}{\sigma_2^2}} \right) \tag{9}
\]

where \( \sigma_1 \) is larger than \( \sigma_2 \), typically about four times. The motivation for this is only that this function has a qualitatively desirable shape, it gives high weights to a ring of points close, but not too close, to the junction point. Any type of function with these and similar properties will do.

Having detected a peak in a histogram, we can use the direction it is representing as a predicted direction in Eq. (7) and approximate the tangent of the curve in the same way as described in Section 3.2.

3.4. Finding the Width of an Edge

Following the notation of Zhang and Bergholm (1993), a diffuse step edge can be modelled by,

\[
E(\sigma, Z_0) = H \Phi \left( \frac{x}{\sqrt{Z_0^2 + \sigma^2}} \right) \tag{10}
\]

where \( H \) models the contrast of the edge, \( Z_0 \) the “true” diffuseness of the edge, \( \sigma \) the extra diffuseness introduced by known smoothing operations and \( \Phi(\cdot) \) is the normal distribution function. They show how \( Z_0 \) and \( H \) can be estimated by controlled variation of \( \sigma \) and least squares fitting. However their method involves edge detection and linking across scale, which could not be called direct. Therefore we compute it differently even though the model is the same. However, since the model is based on a normal distribution function we know that it will be fully determined by its first and second moments. Having made a localization as described above, the covariance matrix can be calculated and the width can be measured from it. That is

\[
C(P(q)) = \begin{pmatrix} c_{xx} & c_{xy} \\ c_{yx} & c_{yy} \end{pmatrix} \tag{11}
\]

where, for instance \( c_{xx} \), is computed by

\[
c_{xx} = \frac{1}{P_{tot}} \sum_{x \in S} \sum_{y \in S} (x - \mu_x)^2 P(x, y) \tag{12}
\]

\[
\mu_x = \frac{1}{P_{tot}} \sum_{x \in S} \sum_{y \in S} x P(x, y) \tag{13}
\]

\[
P_{tot} = \sum_{x \in S} \sum_{y \in S} P(x, y)
\]

The width of the edge is represented by the smallest eigenvalue of \( C \), that is \( Z_0 = \sqrt{\min \lambda(C)} \). Furthermore, the direction of the eigenvectors can be used to check the direction of the fitted line.

3.5. Classification: Straight or Curved

With the method for finding curves presented in this chapter the number of measurements along a curve will be few, but more important is the question if the curve segment is long enough compared to its scale and curvature to characterize its shape. If this is the case a simple criterion can be used. The philosophy is to consider the curve as straight until there is strong evidence that it is curved. A straight segment may turn out to be a part of a curve viewed too closely or in an accidental view. It is required that there should be at least two measurements of direction including the initial estimate for the straight/curved test to be applied.

The criterion used is to threshold on the integrated signed area divided by the length of the curve. It has been used by other authors as well. For instance Bengtsson and Eklundh (1991) used it in multi-scale analysis of shapes of polygonal curves, together with a condition that the absolute deviation of a point from the line between the endpoints should be limited. The integrated signed area is easily computed and has good discriminatory power, see figures 3 and 4. If the curve is represented as a polygon \( p_1, p_2, \ldots, p_n \), then the area is given by the well-known formula:

\[
A = \frac{1}{2} (x_1 y_2 + x_2 y_3 + \cdots + x_{n-1} y_n + x_n y_1 - y_1 x_2 - \cdots - y_{n-1} x_n - y_n x_1) \tag{14}
\]

3.6. Relating Structure Between Fixations

When changing fixation from one point to another, it is important to relate previously derived information
Scheme which yields the shift between two images. This shift is then applied to all previous coordinates in order to perform the transformation of data into the new image.

In this coarse-to-fine correlation strategy an approximate initial shift is given by the approximately known relationship between camera rotations and image shift. At the coarsest level in a Gaussian pyramid, the best normalized correlation match is searched for in an area given by an estimate of the maximum error in the initial shift (in our case about 20 percent). The shift corresponding to this best match is then propagated down through the finer levels of the pyramid, each level correcting the coarser estimate achieved at the previous coarser level. Finally the resulting shift between the two images is given from the finest level with pixel accuracy.

Regarding the image transformation due to the saccade as a translation of the image is of course an approximation and how good this approximation is will depend on the field of view and the amount of rotation that is performed. Using a pinhole camera model and Fick's model for camera rotation around the lens center, see e.g. (Pahlavan, 1993), we can theoretically calculate what kind of errors we can expect. Since it is hard to see what is going on in general from the derived equations, we will look at an example for some specific values of the angles, which are typical for the experiments performed throughout this paper.

A natural constraint on the angles can be foreseen. We would not expect the rotation between fixations to be larger than half the field of view since this would cause the next fixation point to lie outside the image. If this should be the case anyway, we should reside to a lower resolution (changing the focal length), such that the next fixation point lies inside the field of view. This means that the two images, before and after refixa-
tion, will overlap with at least half the image size in both directions. In the plots of the theoretical prediction of the motion between fixations, the rotations are always half the field of view or none at all. The surface shows how the translation varies over the overlapping region. What is done when using the correlation is to assume that the slightly curved surface is planar and parallel to the $u$-$v$-plane, the errors made because of this is exemplified theoretically in figure 5. If a smaller rotation is used with the same field of view the deviations from a plane will be smaller.

As was mentioned earlier, we want to be able to generalize the ideas in this paper to dynamic scenes.
as well. What we picture is that as a point is being fixated, the current model of the object is continuously updated according to the current image of the moving object. This tracking scheme should then account for rigid motion and possibly also deformations of the object. These ideas have not yet been implemented. However, that brings the relevance of the correlation technique into perspective. This should not function alone, but together with some method for tracking the object according to the current model. That is, an image is captured just before and just after the saccade and the correlation provides the corresponding translation, then the model tracker continues at the new fixation point. The saccades are regarded as happening in a very short time period, such that very small changes will occur during that period except for the changes due to the saccade.

In figure 6 we see what happens in a real case taken from one of the experiments. The figure shows the two images from two fixations overlayed, where the overlapping regions are averaged.

3.6.1. Accommodation. For finding the accommodation distance to the edges of a junction, it is possible to vary the accommodation setting and calculate a sharpness measure for a small region at each edge and determine the accommodation distance as the distance that maximizes the sharpness measure.

The sharpness measure used is the Tenengrad criterion (Tenenbaum, 1970), that is

$$\max_d \sum \sum \sum \| \nabla I(x, y, d) \|^2$$  \hspace{1cm} (15)

The accommodation distance is computed at each measurement of the curve.
3.7. Grouping Juncions

A classified junction with its curved or straight legs needs to be incorporated into already detected structure or a model to form a coherent description of a scene or an object. One can imagine numerous strategies, active, reactive or even non-active, for arriving at such a description, of varying complexity and level of abstraction. However, it is not the aim of this paper to discuss such strategies in general. On the other hand it is of importance to show how useful and powerful the suggested approach is for example in providing cues to object recognition. In order to illustrate this, we have hardwired an active strategy with a straightforward and simple method to join classified junctions through their legs as they are detected during exploration of an object. The method incorporates the transformation of previously detected structure into the image at the new fixation point where the most recent candidate junctions are classified, after which these junctions are incorporated into the structure. The locations of legs, which have not been connected, form a basis for selecting new fixation points, in order to connect these legs to other junctions. Finally, the description of an object is completed when no more unconnected legs are found. On the way to the final result, partially complete descriptions are constructed as each fixation and classification occurs.

When the process starts, there is only one classified junction. None of the legs can thus be connected and all legs are considered to be "loose ends", and a new fixation is immediately initiated at one of these ends. The junction classifier has now to verify that the new fixation point really has a connection to the leg which initiated the fixation. If the classifier was unable to verify this the process will refixate at another "loose end". When instead a verification is made, the legs of this new classified junction have to be checked for matches with the "loose ends" incorporated in the description so far. If such a match is found a connection is established and this "loose end" is removed. The legs to which there are no matches, are added to the "loose ends". The process can now continue with the next "loose end", and finishes if there are no more.

T-junctions require special care. If the occluding leg is verified as belonging to the "loose end" that initiated the fixation, the leg is incorporated into this "loose end" which then stays "loose". The occluded leg can be discarded as belonging to another object and thrown away. If on the other hand the occluded leg is verified as belonging to the structure, this leg is also included into the "loose end" but with the difference that this end is removed from the "loose ends", and the occluding leg can be thrown away.\[13\]

4. Experiments and Results

We will here present experiments done in an active environment, that is with the KTH Head-Eye system, see figure 7, for a full detailed description see (Pahlavan, 1993). The setup has been a scene with some simple objects placed about a meter or so from the camera. The fixations have been done with "eye" movements only. This means that since the head rotates its "eyes" around the lens centers, the structures under study will only undergo translation in the image (see Section 3.6 for a discussion). This is important since this makes it possible to transform the data from one fixation to the next using the simple strategy described in Section 3.6.

4.1. Experimental Methodology

The methodology used in the experiments are:

1. Fixate on one object
   (a) Take an overview image of the scene.
   (b) Find a set of junction candidates in this image.
   (c) Make a simple local junction classification and select an $L$, $Y$- or $T$-junction.

2. Foveate, that is fixate and increase the resolution\[14\] on the object under study.

Figure 7. The KTH Head-Eye system was used for performing the experiments. The head-eye system consists of two cameras mounted on a neck and has a total of 13 degrees of freedom. It allows for computer-controlled positioning, zoom and focus of both the cameras independently of each other.
3. For each fixation on this object
   (a) Calculate the transformation from the last fixation.
   (b) Improve localization of the junction candidate.
   (c) Find the dominant directions, establish type of curves and their extent.
   (d) Connect the curves going out from the junction or if a connection cannot be established put them into the list of “loose ends”.
   (e) Choose a new fixation at the end of one the “loose ends” of this junction. If no more “loose ends”, end.
4. Choose new object.

4.2. Grouping Experiments

In figure 8 an overview of a scene with some simple man-made objects is shown, with a spatial sampling per degree of visual angle comparable to what standard camera systems would give. We will now go through the exploration of three objects in this scene—the cube, the truncated cone and the block behind the cone, figures 9–13.

The exploration of the cube starts with a fixation on a strong junction candidate, at an increased resolution. A classification is performed and result is shown in figure 9a.

Refixating at the end of one of the “loose ends” of the cube, as shown in the figure 9(b). When this junction is classified, a match is found with one of the legs from the last junction, which is indicated by a solid line. Continuing with the next fixation a T-junction is encountered (9c). The classification is, in this case, quite clear from geometric information, but this can be supported for instance by using focus information, see figure 10. The difference between the max positions gives this directly. The graphs are calculated from the T-junctions at the top of the cube and the right vertical edge of it. The accommodation measures are calculated at the two windows at the two junctions. One window is placed at the occluding edge and one window at the occluded edge. The window sizes and positions are based on the information obtained in the junction classification.

The search continues all the way round and back to the first fixation on the top face (9d, e). One leg could not be found due to too low contrast between the front faces of the cube (9d). The “loose end” of the first fixation (9a) initiates a refixation at the bottom left corner of the cube (9f). At the fixation on the bottom front corner, the strongest junction candidate is not the corner but a point on the lower edge. This means that this fixation produces two classified junctions to be connected into the group. The first is classified as an edge and just prolongs the edge whereas the second, an L-junction, becomes a new node in the group (9g).

The next object to attend to is the truncated cone, which is a plastic coffee cup turned upside-down. Notably, this object is initially only partially within the field of view. The strongest junction candidate in this case is at the T-junction on the top of the cone, as shown in figure 11(a). In the same manner as before the process refixates at the end of one the legs, the one going to the right. At this point an L-junction with one curved and one straight segment is found (11b). This corner will have two “loose ends”, one merged with the occluding edge of the T-junction. The process will pick one leg, in this case the curved, and fixate at the end of it (11c). A match for the other curved leg of the fixation is not found since only an L-junction was obtained at the other end. The new fixation will be at the end of the straight curve going down on the left side of the cone (11d). Since the curve passes out of the field of view, the default behavior is to restart the junction classification at the new fixation point. This gives a prolonged edge where, at the end, a junction point is found (11e). The same problem occurs at the bottom curve boundary (11f) where a similar result is obtained. Continuing in the same manner the process succeeds to find its way round (11g).

This is also an example of an object in the background—a toy block. Starting at the T-junction at the top of the truncated cone, see figure 12a. In this case, since the front object has been investigated, the grouping process can be instructed to choose the

---

*Figure 8.* The left picture shows an overview of the scene and the 10 most important interest points overlayed on it. Displayed on the right is an edge image produced with the Canny-Deriche edge detector.
Figure 9. The fixation sequence for the cube. The top left corner shows the initial fixation. Then the processing has continued as the pictures are ordered—left to right, top to bottom. Solid curves indicate a segment matched to two junctions, dashed curves are "loose ends" and dotted curves indicate structures which have been classified as belonging to another object. The left pictures in (a) to (h) show the individual classifications and the right pictures the accumulated groupings up to and including the classification to the left.

occluded leg. Even though this boundary is physically curved, it is too short to be labelled as curved. As mentioned before, the principle is to classify a segment as straight until there is strong evidence that it is curved. The search around the top face succeeds (12(b–f)). As before, one leg is missing due to too low contrast between the front faces. Continuing down the vertical visible edge on the right (12(g)). At the junction shown in figure (12(h)), some problems occur and the grouping procedure cannot pass this point with the default
parameter setting. Passing this obstacle (12(i-j)), the grouping becomes complete. This example shows that partly occluded objects can be handled as well.

5. Conclusions

The purpose of this work is to demonstrate how active and selective fixation can be applied to extract meaningful features in a scene in machine vision. Our approach begins by potential interest in the scene, followed by a series of fixations and saccades. As in biology fixation here refers to a mechanical movement of the eyes (cameras). Hence, information is actively extracted from the scene during the process. The mechanism is implemented on a head-eye system and makes relative depth information available through both accommodative and binocular disparities. Saccade for high resolution over a limited field of view is obtained by zooming, simulating a foveated sensor.

Vision is a task-oriented and goal-directed process. Active strategies for where to look next and to decide what information to use, should therefore be considered in view of what the underlying task is. To be able to investigate the role of selective fixations we have in this work the task as that of recognition of man-made objects.

We contend that in an active, continuously operating vision system it is not reasonable to base recognition on complete surface or scene reconstruction. This argument has also been forwarded by e.g. (Crowley and Christensen, 1995). Hence, we propose that such a system should rely on information acquired by selective fixations and by integrating different cues at these fixation points.

In contrast to dominating techniques our framework does not depend on finding, tracing or linking edges. Nevertheless, it is based on information from retinotopic processing of a general nature, like lower-order directional derivatives. The objects considered are characterized by their shape. Essential information is then provided by image junctions and their types. The work reported here aims at showing how such information can be derived from real scenes using the proposed framework.

As argued by Biederman and Malik crucial information is in this context given by the straightforwardness or curvedness of the edges meeting. Such properties can be derived from the locations of the attention points which are predicted by the directions of maximum response at the junction and adaptively modified at each attention step. Our anthropomorphic head-eye system will provide the mentioned depth cues. They are used to disambiguate certain accidental configurations.

Generally, we have shown that the selection of attention, driven by the specific task, can generate coherent and meaningful groupings of features in the scene, in this case junctions. Notably, the ubiquitous problem of figure-ground segmentation is efficiently handled in our scenario. We believe that the so abstracted sets
of junctions are directly useful for model indexing. In particular, since they contain few spurious features they avoid the risk of combinatorial explosion.

The approach has been integrated into a working active system using the KTH Head-Eye system.

Although we have recognition as our background task, the reported work does not actually deal with model indexing or matching. It is worth observing that our process can, when coupled to such techniques, already at an early stage say when two or three junctions
Figure 12. The result after processing a toy-block partly occluded by the truncated cone. (a) The grouping start with a T-junction. The occluded leg is too short for being classified other than straight. (b)-(g) The search finds its way round the top face. Similarly as before the inner boundary cannot be found due to too low contrast. (h) Some problems with the structure in background was encountered. The search did not succeed to pass this point with default parameters. (i)-(j) The grouping finds all possible matches. This example shows that the grouping can handle also partly occluded objects.

Figure 13. The groupings found in the examples shown in figure 9 (the cube), figure 11 (the truncated cone) and figure 12 (the occluded toy block).
have been grouped. Such a subgroup would contain enough information for narrowing-down the possible interpretations to just a few, even in a quite large database. The continued grouping process could then be guided by the knowledge of objects given by the models, and the interpretation could, when a new junction was found and classified, be updated and in this way result in cooperation between the modules. Recent work on indexing is e.g. (Wilkes and Tsotsos, 1993a).

The cues to depth are very important for many reasons. In the presented work we have only used them in cases which otherwise could not be resolved, for instance at accidental views. This means that having reliable relative depth information, is of crucial importance. The different cues to depth have complementary qualities. Binocular disparities give accurate measures, but structure along the epipolar line will be unreliable. Accommodation on the other hand does not have this limitation, but the accuracy could sometimes be too low to give any conclusive information. Work on incorporating both these cues is, therefore, needed to integrate the accuracy of binocular disparities with the stability of accommodation.

In creating a coherent interpretation of a whole scene it is not reasonable for a continuously operating system to remember every little detail it has ever seen. This is certainly true for humans, because it is always possible to take a second look. We here get an indication of problems on the coupling between foveal and peripheral vision, but also on how to maintain a model stable of the world. When revisiting an object to recapture information about something that has been forgotten or obtain something that was not extracted at first glance, at least its approximate location should be available, so a complete search of the scene would not be necessary. What this discussion shows is that active vision, even though successfully applied in our framework, leaves many interesting questions to be answered.

\[
\Sigma_{k/k} = \Sigma_{k/k-1} - \Sigma_{k/k-1} H_k (H_k^T \Sigma_{k/k-1} H_k + R_k)^{-1} \times H_k^T \Sigma_{k/k-1}
\]
\[
\Sigma_{k+1/k} = F_k \Sigma_{k/k} F_k^T + G_k Q_k G_k^T
\]

as described in e.g. (Anderson and Moore, 1979). \( z_k \) is a vector containing the entities that are measured in measurement \( k \), the covariance matrix of this measurement is \( R_k \). This yields an estimate, \( \hat{x}_{k/k} \), of the state at \( k \), taking into account the last measurement \( z_k \). This in turn gives a prediction, \( \hat{x}_{k+1/k} \), of what the state should be for the measurement \( k + 1 \) with corresponding covariance matrix \( \Sigma_{k+1/k} \). This is based on the signal model

\[
x_{k+1} = F_k x_k + G_k w_k
\]
\[
z_k = H_k^T \hat{x}_k + v_k
\]
\[
E[ww^T] = Q_k \delta_{kl}, E[v_k v_k^T] = R_k \delta_{kl}
\]

In our case we wish to determine a curve in the plane. The used model of the curve to determine is a parameterized function of the type

\[
p(t) = (x(t), y(t))^T
\]

represented by its Taylor expansion up to the second order. Measurable quantities of this curve which we will use here are positions on the curve and the corresponding tangential direction, contrast and scale. This gives \( x_k, F_k, G_k \) and \( H_k^T \) in the signal model above

\[
x_k = \begin{pmatrix} p_k \\ v_k \\ a_k \\ c_k \\ s_k \end{pmatrix}
\]
\[
F_k = \begin{pmatrix} 1 & 0 & \frac{1}{2} \Delta t L & 0 & 0 \\ 0 & 1 & \Delta t L & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, G_k = I
\]
\[
z_k = \begin{pmatrix} p_k \\ v_k \\ a_k \\ c_k \\ s_k \end{pmatrix}, H_k^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}
\]

Here \( p_k \) in \( x_k \) keep track of the 2D position of the curve, \( v_k \) is the tangential direction of the curve, \( a_k \) the change.

A. Kalman Filter Details

The filter used for the estimation of the different parameters in Section 3.2 is a Kalman filter divided into a prediction and a measurement part of the form

\[
\hat{x}_{k/k} = \hat{x}_{k/k-1} + \Sigma_{k/k-1} H_k (H_k^T \Sigma_{k/k-1} H_k + R_k)^{-1} \times (z_k - H_k^T \hat{x}_{k/k-1})
\]
\[
\hat{x}_{k+1/k} = F_k \hat{x}_{k/k}
\]
in the tangent direction, while \(c_k\) and \(s_k\) are the contrast and the scale of the intensity transition at the curve. Other measurable parameters, such as relative depth, diffuseness, grey-level curvature etc., can also easily be incorporated into the filter. The contrast and the scale are assumed to be approximately constant.

The Kalman filter is initialized with

\[
\hat{x}_{0/-1} = \begin{pmatrix} p_0 \\ q_0 \\ 0 \\ c_0 \end{pmatrix},
\]

\[
\Sigma_{0/-1} = \begin{pmatrix} \Sigma^p_0 & 0 & 0 & 0 \\ 0 & \Sigma^q_0 & 0 & 0 \\ 0 & 0 & \Sigma^c_0 & 0 \\ 0 & 0 & 0 & \sigma^c_0 \end{pmatrix}
\]

(26)

where the components of \(\hat{x}_{0/-1}\) are taken directly from the estimates of the initial directions described in Section 3.3, whereas the covariances are just set to high values to give high weights to the new measurements of points on the curve in the beginning. Each new measurement then yields an estimate of the current state, \(\hat{x}_{k/k} = (\hat{p}_{k/k}^T \ \hat{q}_{k/k}^T \ \hat{a}_{k/k}^T \ \hat{s}_{k/k}^T)\), through Eq. (17).

The \(\Delta_{k}\) in \(F_k\) is not kept constant along the curve but is instead allowed to vary as to control how far along the curve we should make the next estimation. As described in Section 3.3 the predicted position then gives the search area in which the actual position of the curve is searched for. \(\Delta_{k}\) starts with a low value and is increased at each step, unless the process fails to find the curve at the new location. In that case the step length is decreased, as illustrated by figure 2. The update is done according to

\[
\Delta_{k} = \Delta_{k-1} + \beta_k \cdot \hat{s}_{k/k}
\]

(27)

where \(\beta_k = \beta_0\). The scale, \(\hat{s}_{k/k}\), is coming from the update of the filter, Eq. (17), and \(\beta_0\) and \(\Delta_0\) are constants. It should be noted here that the filter is only updated if the curve was found using this \(\Delta_{k}\) in the prediction. If this is not the case the prediction is made with a smaller \(\Delta_{k}\), also given by Eq. (27), but with \(\beta_k = -n\beta_0\). Increasing \(n\) by one for each failure to find the curve from the new estimation. The filter stops if \(\Delta_{k} < \Delta_0\).

As described in Section 3.2, the measurement of the location and tangential direction of the curve will be searched for in a sector formed by an uncertainty angle and the predicted direction, at the prediction obtained from the prediction step of the Kalman filter.

B. Additional Experiments

We present in this appendix two experiments illuminating some aspects of our approach. The first one stresses the importance of depth information and the other shows an example where the background is textured. For a larger collection of experimental results see (Brunnström, 1993).

B.1. Accidental Alignments

This experiment will illustrate how important it is to be able to support a classification with depth information. In fact, in some situations it is the only way to arrive at the correct conclusion. For instance if an edge in the background is passing right through or very close to a junction point, as at the junction on the top of the cube in figure 14, it is impossible to find out locally from geometrical considerations if this edge belongs to the object or not. With depth information available, however, it is a simple task to make such correction.

In this example we will demonstrate three situations where this comes in natural. First there is an accidental alignment with an \(L\)-junction of a cube, see figure 15, forming a \(Y\)-junction in the image. Analyzing the depth measurements from accommodation at the object and at the accidentally aligned edge the depth difference

Figure 14. An overview of the scene.
Figure 15. The fixation on a junction where an edge in the background is accidentally aligned with corner of the object. The top row shows the classification based on geometry alone and the bottom row if depth information is incorporated as well. The accumulated structure is shown to the left and the junction classification to the right.

Figure 16. The different depth measurements obtained by accommodation for a "false" Y-junction (upper right in figure 15). The depth measures along the y-axis are not absolute, but higher values are further away than lower values. The numbers along the x-axis are the ordering numbers of measurements going out from the junction point, i.e. lower values are closer to the junction. They correspond to focus measurements made during the process described in figure 2 in which the measurement at a fixation point is incorporated using the method described in Appendix A. The upper line shows how the depth varies along the occluded edge while the lower line shows the variation along the left edge. It is clearly seen that these edges lie on different depths, and also that different measurements along the same edge are consistent.
Figure 17. An attended T-junction. The accumulated structure is shown to the left and the junction classification to the right.

Figure 18. The different depth measurements obtained by accommodation for a T-junction (right in figure 17). The depth measures along the y-axis are not absolute, but higher values are further away than lower values. The numbers along the x-axis are the ordering numbers of measurements along the edge. They correspond to focus measurements made during the process described in figure 2 in which the measurement at a fixation point is incorporated using the method described in Appendix A. The upper line shows how the depth varies along the occluded edge while the lower line shows the variation along the occluding edge. Measurement 0 is closest to the junction for the occluded curve. As for the measurements along the occluding edge, the junction lies between measurement 3 and 4, and the measurements go outward from there, i.e. 3 is closer to the junction than 2 and similarly 4 is closer than 5 etc. Some problems with consistency can be seen along the occluding edge although all measurements clearly indicate that this edge lies closer. Together with the strong geometrical information, this supports that there is indeed a depth discontinuity.

clear, as shown in figure 16. The background edge has been removed from the junction grouping to illustrate this.

The next situation is at a T-junction, see figures 17 and 18. The depth information here gives support to the geometrical information, which is very strong in this case.

Finally, at the bottom right of the cube there is a shadow aligned with the edge forming a T, see figure 19. It is a weak geometrical T-configuration and the depth difference is also small, thus the overall support for the classification is small, as plotted in figure 20.

Some general comments on the accommodation graphs (figures 15, 17 and 19). During the experiments an initial accommodation was done for each refixation at a junction candidate. An interval of images with different accommodation was then taken around this initial accommodation. This interval corresponds to the range 0 to 10, in which 5 is the initial accommodation. This
means that the values for the relative depth cannot be compared between the graphs. To get the absolute depth this would require knowledge of the focal length used during accommodation, the initial focusing position and the focusing interval. Of course a calibration of the relation between focal length, focus position and absolute depth has to be made. However, no calibration was done prior to these experiments.

B.2. Processing of a Cube Against Textured Background

We now consider a cube against a textured background, see figures 21–23. More than half of the visible junctions and about two thirds of the visible edges have been found, but the process has problems with finding good junction candidates in the textured region as can
Figure 21. To the left is shown an overview of the scene and to the right the line segments obtained by a standard edge detector followed by edge tracing and line fitting. It illustrates the complexity of grouping strategies based on such data. Note, that this data has not been used in this algorithm.

Figure 22. The first three junctions attended to in this scene. The accumulated structure is shown to the left and the junction classification to the right.

Figure 23. The forth to the sixth junction attended to in this scene. The accumulated structure is shown to the left and the junction classification to the right.
be expected. One reason for this is that the default parameter setting was chosen for finding isolated junction candidates and when this was not fulfilled, no candidate was found. Note that at the edges where a direction prediction has been obtained, most of the edge-curves are found, even against the textured background. We support this by supplying what the classification would have been if the junction candidates had been found, see figure 24.

B.3. Evaluation of the Classification in the Experiments

To evaluate the classifications, all the junctions and the classified edge-curves have been counted and their correctness have been subjectively determined in two ways; one is based on the image data and the other one is based on the knowledge of the underlying object. For instance, considering a corner resting on a table with a shadow attached to it clearly visible, then if all the object edges and the shadow have been found, the classification is considered as correct based on image data, but incorrect based on the knowledge about the underlying object. In the junction case the rate of correctness is 88% and 75% respectively and in the edge-curve case the rates are 98% and 95% respectively, see figure 25.

C. Experiment with Controlled Added Noise

Real imagery normally contains a great deal of noise. To model this realistically is very complicated, furthermore it is often very difficult to control its level. In order to demonstrate the robustness of the junction classification against noise we have added Normal distributed noise of different magnitude to an image and performed classifications of five junctions for these cases, see figures 26–30. The standard deviations of the added noise have been 0 (no noise added), 5, 10, 20, and 50. These examples show that the classification procedure has strong resistance to noise and it really does not break down until the signal-to-noise ratio is about 50–100%, see figure 30.

Acknowledgments

The reported work forms part of the ESPRIT-project, Vision-as-Process II BR 7108 and a project in the basic program of Computer Vision of the Swedish National Board for Industrial and Technical Development, NUTEK. This support is gratefully acknowledged.

Notes

1. Biederman’s experiments involve no eye movements, which is different from our situation. Our focus is on what the cues are.
2. Active acquisition of new image data with locally highly increased resolution.
3. The different edge labels such as “convex”, “concave” and “limb” have not been considered, and “occluding edge” has only been used in the case of T-junctions.
4. We will relax this principle and only perform refixation when attended structure is not visible anymore or too peripheral and just move the attention point, but we will discuss the method as if a fixation is really performed at each step.

<table>
<thead>
<tr>
<th>Type</th>
<th>T</th>
<th>Cim</th>
<th>Obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>junction</td>
<td>60</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>curve</td>
<td>133</td>
<td>131</td>
<td>126</td>
</tr>
</tbody>
</table>
Figure 26. Classification of five junctions with added noise with Normal distribution (a)-(e). In this case no noise was added. The graph in (f) is showing the pixel values along a fixed y-coordinate and the location of this row is indicated by the small arrow to the left.

Figure 27. Classification of five junctions with added noise with Normal distribution (a)-(e). In this case the noise distribution was $N(0, 5)$. The graph in (f) is showing the pixel values along a fixed y-coordinate and the location of this row is indicated by the small arrow to the left.
Figure 28. Classification of five junctions with added noise with Normal distribution (a)–(e). In this case the noise distribution was $N(0, 10)$. The graph in (f) is showing the pixel values along a fixed y-coordinate and the location of this row is indicated by the small arrow to the left.

Figure 29. Classification of five junctions with added noise with Normal distribution (a)–(e). In this case the noise distribution was $N(0, 20)$. The graph in (f) is showing the pixel values along a fixed y-coordinate and the location of this row is indicated by the small arrow to the left.
Figure 30. Classification of five junctions with added noise with Normal distribution (a)-(e). In this case the noise distribution was $N(0, 50)$. The graph in (f) is showing the pixel values along a fixed y-coordinate and the location of this row is indicated by the small arrow to the left.

5. By this we mean one of the edges meeting at the junction.
6. The notation used here is that "indicates a prediction and "a measurement."
7. The choice of the radius has not been automated, during the experiments it has been set to a constant.
8. The gradient is calculated using central differences in x and y directions on a grey-level pyramid computed with binomial filters, but the method presented here does not rely on a particular way of obtaining it.
9. Exchange $S_{i+1}$ with the filter prediction $S_{i/k+1}$.
10. Note here that we are doing local measurements at a high resolution, therefore we assume that the intensity along an edge does not vary much, except at a few locations. This is of course a simplified model of the world which could be subject to revision. The same assumption is the basis of the histogram classification of junctions.
11. In a foveated system this would automatically happen since we have a decrease in resolution towards the periphery and hence as the new fixation point lies further out.
12. The matching criterion used here is one which joins the polygons of two legs and calculates the area surrounded by this joined polygon. This area is divided by the square of the longest distance between two points in the polygon, $A/P^2$, which is thresholded for match. In the examples the thresholds were 0.02 for straight legs and 0.06 for curved.
13. In these experiments we have always thrown away one of the "legs" in a T-junction. On more articulated objects with self occlusion, the two legs may in fact belong to the same object. The fact remains that there is a discontinuity, and whether this discontinuity lies on the same object or not cannot be decided locally, it has to be decided later. We have chosen here to simply keep what can immediately be incorporated into the already existing structure and throw away the rest. A straightforward extension would be to instead keep the rest, and later on throw it away if there is no other connection to it.
14. The principle is to get a resolution better matching the human foveal vision.

15. Does not have to involve moving the camera and acquiring new data in our experiments.
16. At this stage all of the legs of the junction is "loose ends".
17. The dashed curves will indicate "loose ends", dotted curves are considered to belong to another grouping and will not be considered further, and the solid curves are those that have been connected between two junctions.
18. From this example and forward we are only performing fixation when necessary and moving the attention point otherwise.
19. The scale here is the scale of the intensity change across the edge and the contrast as function of the gradient magnitude.
20. No automatic decision has been implemented yet for handling this.
21. Since the junction grouping process disconnect a grouping from the surrounding at T-junctions, the evidence for a T-junction should be strong for allowing a disconnection otherwise it should be kept in the grouping until more information has been obtained.
References


