Active Fixation for Scene Exploration*

Kjell Brunnström, Jan-Olof Eklundh and Tomas Uhlin

Computational Vision and Active Perception Laboratory (CVAP)**
Royal Institute of Technology (KTH), Stockholm, Sweden

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 by its nature active and tied to the behaviors of a seeing agent. 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) .

Most of this work has been devoted to architectural problems and to the problem of controlling gaze in real-time. The far more general problem of actually achieving active and purposive vision has been given much less attention, even though especially Ballard (1989) and Aloimonos *et al.* (1990) have argued for such approaches.

Exceptions exist, like in the work of Bajscy and Campos (1992), which discuss a general framework of control in what is called the "where to look next" problem, in Rimey and Brown (1992), who apply a probabilistic framework to address the same problem. In both cases it is studied from a high-level perspective. There are still many open questions concerning how primary processes of fixation and gaze holding relate to the high level tasks and to the intentions and motivations of the observer. One such question concerns if and how the active approach can support recognition by providing mechanisms for addressing the figure-ground problems. It is well-known that most techniques for recognition and model indexing in machine vision suffer from combinatorial problems. The main reason for this seems to be

*The support from the Swedish National Board for Industrial and Technical Development, NUTEK, is gratefully acknowledged.
**Address: NADA, KTH, S-100 44 Stockholm, Sweden
Email: kjellb@bion.kth.se, joe@bion.kth.se, tomas@bion.kth.se

BMVC 1993 doi:10.5244/C.7.6

that it is difficult to group features into groups that are meaningful *in the scene*, a difficulty that is hardly noticeable for a human observer in a real environment. In this paper we consider this problem in a limited context using an active approach.

As mentioned above it is reasonable to consider the problem in relation to some task or context, so that the system can be provided with the necessary "background" knowledge. The underlying task here 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. Furthermore, forming groupings of such junctions separated from the surrounding at depth discontinuities i.e. T-junctions in monocular images, would provide useful input for matching with parameterized models. 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) and Westelius *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. Furthermore, the approach does not require that all the interesting parts of the scene are in the field of view initially.

2 Background

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. Biederman (1985)(1987), has shown that humans use qualitative cues of a similar nature, including edges classified as straight or curved, to do rapid model indexing[1].

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 multi-local 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 junction and edge types. The only information used is what one gets 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. Since fixation in active vision means not only selecting a region of interest in the image but rather a region of interest in the scene domain, depth cues from binocular disparities or accommodation are also available. We will show how they can be integrated in the classification scheme as well.

The outline of the remaining paper is as follows. We first describe the computational principles of our approach, including the integration of additional cues. We then present some experimental results and end with a discussion.

[1] Incidentally, Biederman consider cases where no eye-movements occur. However, our reference is only aimed at motivating what cues to use.

3 Computational Aspects

The underlying task, which we consider, is to be able to do recognition. At low level, some features, suitable for the task, have to be extracted. Here, we consider scenes with manufactured objects for which classified junctions in connected groups are suitable features. In this section we will discuss how these can be obtained in an active scenario.

To cue recognition the system explores junctions. The goal is to obtain clusters of connected junctions, with connecting curves labelled as straight or curved, separated from the surroundings at depth discontinuities i.e. T-junctions. The exploration can be initiated by starting at a point indicated by an attentional process for instance as discussed in e.g. Brunnström *et al.* (1992). The classification method, establishes the number of curves meeting, their type – straight or curved, and also tries to find the extent of the curves. This is done in a context of an active environment, which means that zoom, focus and stereo can be incorporated in the analysis as well. The clusters of junctions are then obtained by refixating at the end of one leg of the junction and search for a new junction compatible with the already found curve. A new classification is performed at the new fixation and the search is continued in a depth-first manner. The other legs are queued until later. The detected and classified junctions will thus be connected through their matching legs. Some parts of this process will be described in more detail, the strategy to classify curves meeting at the junction, the method for relating structure from one fixation into the next, how relative depth can be incorporated and the used grouping strategy.

3.1 Changing fixation along a curve

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 fixation point in the direction of the expected direction of the edge for a small number of steps.

Initially at the junction an estimation of dominant directions in its immediate surrounding is performed. The method used is presented in Brunnström and Eklundh (1993) and will not be further developed here.

An underlying idea of this method is, that having an estimate of the curve direction at one point, it should be possible to move the fixation point some step and then find the curve again, without having to trace it pixel by pixel. What is needed is some scheme, where measurements of the curve are taken as input and predictions of its location, direction etc. are calculated. A powerful tool for 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, disparity etc. The process can roughly be described by Figure 1(a). 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. For further details see Brunnström *et al.* (1993).

The curves are classified into straight or curved by computing the integrated signed area divided by the length of the curve, as illustrated by Figure 1(b).

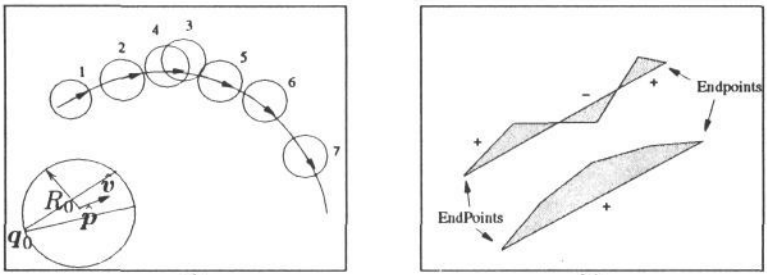

Figure 1: *a) Illustration of the selective fixations along a curve. b) Illustration of the curve classification criteria.*

3.2 Relating structure between fixations

When changing fixation from one point to another it is important to relate previously derived information to the image acquired at the new fixation point. This is done by transforming all the coordinates belonging to the extracted data into the coordinate system given by the new image. In a static scene we could rely on carefully calibrated cameras in order to find the transformation needed using the known rotations of the cameras. If we instead do not wish to rely on careful calibration and also hope to generalize the approach to dynamic scenes, we have to use an image based method to find the transformation. In these examples this was done with normalized correlation in a coarse-to-fine 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.

3.3 Relative Depth

Low level processes would be aided by having depth information available. This does not have to be an exact absolute quantity from a calibrated stereo system, but rather a qualitative relative one, which could discover depth discontinuities. The decisions to make are whether the features under study are at the same depth or not. For instance this would disambiguate accidental alignments and help verifying the classifications of junctions. At T-junctions the case of surface markings has to be ruled out. A junction classified as an L, could in reality be a T, with a part of the occluding edge missing, which is not an uncommon case. Even Y- and ↑-junctions need to be checked whether all the crossing curves can be said to meet at approximately same depth.

There are two fundamentally different ways of obtaining this information. One is based on using both "eyes" in a vergence disparity method and the other one is based on accommodative cues – "depth-from-focus" or "depth-from-defocus". Disparity will at structure normal to the epipolar line and in non-occluded regions, give accurate relative distance measures. The problem is that at the locations where it is needed the most, close to depth discontinuities, it is very hard to get reliably. Accommodation, on the other hand, since it is a monocular cue does not have these problem and the accuracy is enough for our purposes and is the method adopted in these experiments. It has not yet been fully integrated into the classification scheme, but examples how it can be used are shown in Figure 4. However, we plan to incorporate both these two methods in the future, to combine the accuracy of disparity with the robustness of accommodation.

3.4 Grouping junctions

A classified junction with its curved or straight legs need to be incorporated into already detected structure or a model to form a coherent description of a scene or 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, forms 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 it will refixate at another "loose end". When instead a verification is made, the legs of this new classified junction has to be checked for matches[2] 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. Care has to be taken at T-junctions. If the occluding leg is verified as belonging the "loose end" that initiated fixation, the leg is incorporated into this "loose end" which then stay "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.

4 Experiments and Results

We will here present experiments done in an active environment, that is with the KTH Head-Eye system, Pahlavan and Eklundh (1992). 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 with "eye" movements only, thus the structures under study will only undergo translation in the image[3], making it possible to transform

[2] 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/l^2, which is thresholded for match. In the examples the thresholds were 0.02 for straight legs and 0.06 for curved.

[3] Rotations in the image will also be present but these can be neglected since camera rotation is small and thus the resulting image rotation will be small

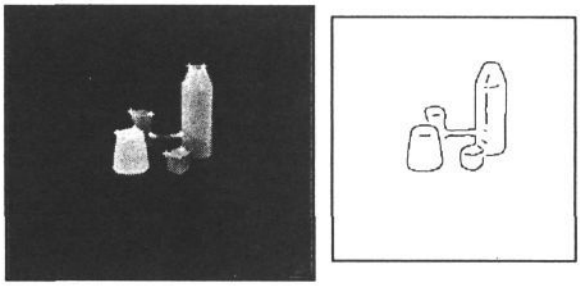

Figure 2: *The left picture show an overview of the scene and the 10 most important interest point overlayed on it. Displayed on the right is an edge image produced with the Canny-Deriche edge detector.*

the data between fixations using the simple strategy described in Section 3.2.

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) Draw one of the strongest as the junction to attend to.
2. Foveate, that is fixate and increase the resolution on the object under study[4].
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 the type of curves meeting at the junction 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 2 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.

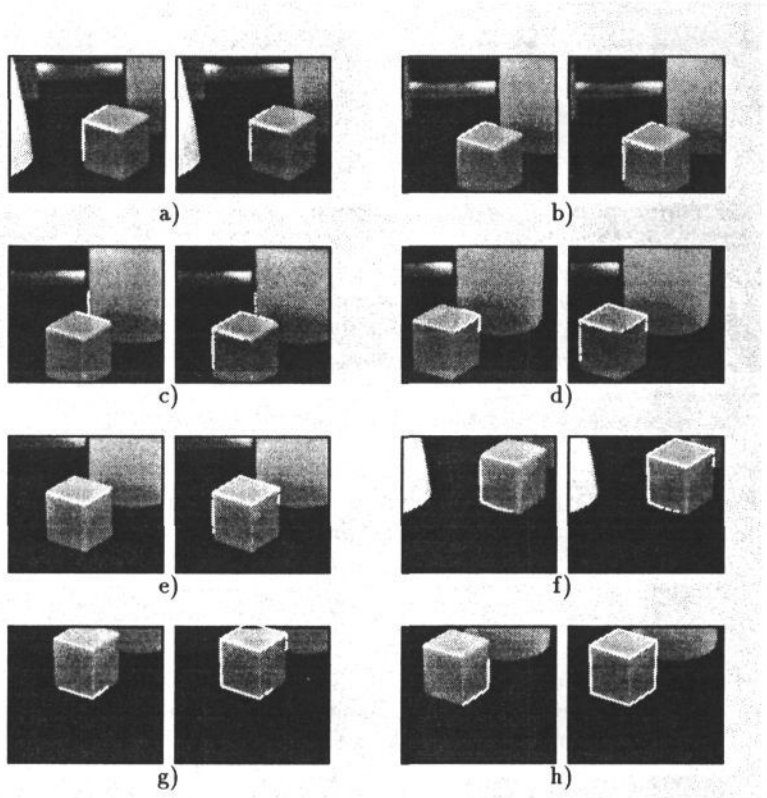

Figure 3: *The fixation sequence for the cube. The top left corner shows the initial fixation. Then the processing have continued as the picture are ordered – left to right, top to bottom. Solid curves indicates a segment matched to two junctions, dashed curves are "loose ends" and dotted curves indicate structure which have been classified as belonging to another object. The left pictures in a) to h) shows the individual classifications and the right pictures the accumulated groupings up to and including the left classification.*

The exploration of the cube starts with a fixation on a strong junction candidate, at an increased resolution. A classification is performed and the result is shown in Figure 3(a).

Refixating at the end of one of the "loose ends"[5], as shown in Figure 3b[6]. 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 fixations a T-junction is encountered (3c). The classification is, in this case, quite clear from geometric information, but this can be supported for instance by using focus information, see Figure 4.

The search continues all the way round and back to the first fixation on the top face (3d,e). One leg could not be found due to too low contrast, between the front faces of the cube (3d). The "loose end" of the first fixation (3a) initiates a

[4] The principle is to get a resolution comparable with to what humans have in the fovea.

[5] At this stage all of the legs of the junction are "loose ends".

[6] 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.

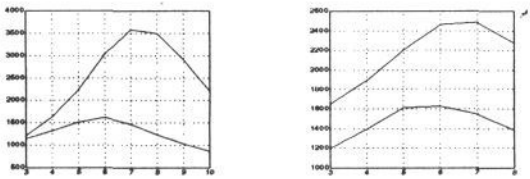

Figure 4: *Focusing measure at different accommodation distances for two T-junctions. The left graph is from the T-junction at the top of the cube and the right graph is from the one on the vertical right edge of the cube. The two curves in the graphs have been calculated at the occluding and occluded edges respectively. The calculation areas are computed from the curves of the found junctions. In both cases the measure attains maximum about one unit of 32 from each other, which corresponds to about 5 cm on 1 m distance. The x-axis represents the focal values, where lower values means closer to the observer and the sharpness measure are plotted along the y-axis.*

refixation at the bottom left corner of the cube (3f). 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 (3g).

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 5a. 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 (5b). This corner will have two "loose ends", one merged with the occluding edge of the T-junction. The process will now give priority to curved leg and fixate at the end of it (5c). 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 (5d). Since the curve passes out of the field of view, the default behaviour is the restart the junction classification at the new fixation point. This gives a prolonged edge where, at the end, a junction point is found (5e). The same problem occurs at the bottom curve boundary (5f) where a similar result is obtained. Continuing in the same manner the process succeeds to find its way round (5g).

Finally we have an example of an object in the background – a toy block. The result of this processing is shown in Figure 6(a).

5 Summary and Discussion

Active visual processes are task-oriented and goal directed. Active strategies for where to look next and to decide what information should be used, should therefore be considered in view of an underlying task. In this paper the task at hand is recognition of man-made objects.

In an active, continuously operating vision system it is not reasonable to base recognition on complete surface or scene reconstruction. This point has been stressed by e.g. Aloimonos, but also follows from efficiency considerations. Instead such a system should rely on information acquired by selective fixations and by

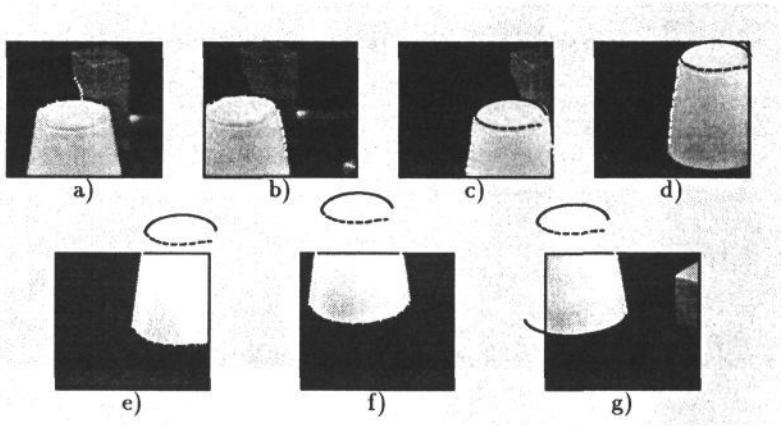

Figure 5: *The grouping process of the cone. a) Shows the initial fixation. This is classified as a T-junction, which means that the occluded curve can be excluded. b) The next fixation is at the right end of the curve. Now an L-junction is matched to curve segment. This gives two "loose ends" and with a priority to investigate the curved segments before straight, the next fixation becomes at the other end of this curve. c)-g) The search continues round the bottom. In this example two fixation are necessary simply because parts of the object are out of the field of view (d and f). The curve segments are marked in the same way as before, but also, the curves classified as straight are drawn in white and the curves classified as curved are drawn in black.*

integrating different cues at these fixation points.

We have explored such an approach, that does not depend on finding, tracing or linking edges, in the context of recognizing objects that can be represented by geons. Essential information is then provided by image junctions and their types, as has been demonstrated by Malik and Biederman.

Figure 6: *a) The result after processing a toy-block partly occluded by the truncated cone. b) The groupings found in the examples shown in Figure 3 (the cube), Figure 5 (the truncated cone) and in (a) of this figure (the occluded toy block).*

We have here and in previous work shown that using information of this type, we can first perform an attentional step, which detects potential junctions, then do a local coarse classification and finally select a few nearby fixation points, that allow us to do a more distinguishing classification of the junctions with regard to their 3-dimensional structure. Important classes, like the categories given by Malik, are given by the straightness or curvedness of the edges meeting. This can be derived from the locations of the fixation points which are predicted by the directions of maximum response at the junction and adaptively modified at each

fixation step. Apart from these geometric and photometric image cues, we can compute binocular disparities or accommodative cues to detect which junctions are T-junctions.

On top of this we have shown that the selection of fixations, driven by the specific task, can generate coherent groupings of features in the scene, in this case junctions. Furthermore, we believe that the abstracted sets of junctions we derive here are directly useful for recognition.

References

[Aloimonos *et al.*, 1987] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. "Active Vision". In *Proceedings First International Conference on Computer Vision*, pages 35–54, 1987.

[Aloimonos, 1990] Y. Aloimonos. "Purposive and Qualitative Active Vision". In *Proc. DARPA Image Understanding Workshop*, pages 816–828, 1990.

[Bajcsy and Campos, 1992] R. Bajcsy and M. Campos. "Active and Exploratory Perception". *CVGIP: Image Understanding*, 56(1):31–40, July 1992.

[Bajcsy, 1985] R. Bajcsy. "Active Perception vs. Passive Perception". In *Proceedings Third IEEE Workshop on Computer Vision*, pages 55–59, Bellair, Oct 1985. IEEE.

[Ballard, 1989] D.H. Ballard. "Animate vision". In *Proceedings 11th IJCAI*, Detroit, 1989.

[Biederman, 1985] I. Biederman. "Human Image Understanding: Recent Research and a Theory". In *Human and Machine Vision II*, pages 13–57. Academic Press, 1985.

[Biederman, 1987] I. Biederman. "Recognition-by-components: A theory of human image understanding". *Psychological Review*, (94):115–147, 1987.

[Brunnström and Eklundh, 1993] K. Brunnström and J.-O. Eklundh. "Active fixation for junction classification". In *Proceedings 5th Inter. Conf. on Computer Analysis of Images and Patterns*, Budapest, Hungary, Sep. 1993.

[Brunnström *et al.*, 1992] K. Brunnström, T. Lindeberg, and J.O. Eklundh. "Active detection and classification of junctions by foveation with a head-eye system guided by the scale-space primal sketch". In G. Sandini, editor, *Proceedings Second European Conference on Computer Vision*, volume 588 of *Lecture Notes in Computer Science*, pages 701–709. Springer-Verlag, May 1992. (Santa Margherita Ligure, Italy).

[Brunnström *et al.*, 1993] K. Brunnström, J.-O. Eklundh, and T. Uhlin. "Active fixation for Scene Exploration". In *Tech. Rep., ISRN KTH/NA/P-93/11-SE, CVAP123*, 1993.

[Culhane and Tsotsos, 1992] S.M. Culhane and J.K. Tsotsos. "An Attentional Prototype for Early Vision". In *Proceedings Second European Conference on Computer Vision*, pages 551–562, Santa Margherita Ligure, Italy, May 1992.

[Koenderink and van Doorn, 1990] J.J. Koenderink and A.J. van Doorn. "Receptive Field Families". *Biological Cybernetics*, 63:291–375, 1990.

[Malik, 1987] J. Malik. "Interpreting Line Drawings of Curved Objects". *International Journal of Computer Vision*, (1):73–104, 1987.

[Pahlavan and Eklundh, 1992] K. Pahlavan and J.O. Eklundh. "A head-eye system — analysis and design". *Computer Vision, Graphics, and Image Processing: Image Understanding*, 56(1):41–56, 1992.

[Pahlavan *et al.*, 1992] K. Pahlavan, T. Uhlin, and J.O. Eklundh. "Integrating Primary Ocular Processes". In *Proceedings Second European Conference on Computer Vision*, pages 526–541, Santa Margherita Ligure, Italy, May 1992.

[Rimey and Brown, 1992] R.D. Rimey and C.M. Brown. "Where to Look Next Using a Bayes Net: Incorporating Geometric Relations". In *Proceedings Second European Conference on Computer Vision*, pages 542–550, Santa Margherita Ligure, Italy, May 1992.

[Ullman, 1987] S. Ullman. "Visual Routines". In W. Richards and S. Ullman, editors, *Image Understanding 1985-86*. AbIex, Norwood, N.J., 1987.

[Waltz, 1975] D. Waltz. "Understanding Line Drawings of Scenes with Shadows". In Winstor P.H., editor, *The Psychology of Computer Vision*. McGraw-Hill, New York, 1975.

[Westelius *et al.*, 1991] C.-J. Westelius, H. Knutsson, and G. H. Granlund. "Focus of Attention Control". In *Proceedings of the 7th Scandinavian Conf. on Image Analysis*, pages 667–674, Aalborg, Denmark, 1991.