Issues in active vision: attention and cue integration/selection

Jan-Olof Eklundh, Peter Nordlund, and Tomas Uhlin
Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computing Science
Royal Institute of Technology, S-100 44 Stockholm, Sweden
joe@nada.kth.se

Abstract

In this paper which is a summary of our talk, we stress a systems approach for research in active vision. We also argue that design and analysis of seeing agents should be accompanied by experiments, requiring implementations, i.e. a constructive approach. In particular, we discuss two issues that we have worked with: use and integration of multiple cues and attention.

1 Introduction

During the past decade computer vision researchers have devoted considerable attention to vision as a guide to action. This work has been performed under the names of Active, Animate, Purposive or Behavioral Vision, see e.g. (Bajcsy, 1985; Aloimonos et al., 1987; Ballard, 1991; Aloimonos, 1990; Nelson, 1991) The perspective is not new: perception and action in conjunction have for a long time been studied within fields such as robotics, AI, and also in the neurosciences, psychology and cognitive science. In fact vision as a means to guide robots was considered already in the 1960's. No doubt, there is a wealth of interesting results on vision and action, see e.g. (Fermüller and Aloimonos, 1995; Eklundh, 1995) for discussions, but as pointed out in the introduction of the first of these two papers there exists no firm theoretical foundation for integrating vision and action. It even seems to be an open question how the problem of devising artificial seeing agents should be approached scientifically at all. Although it may be difficult to answer that question at present, we believe that one can identify some key issues in this context. In our talk, which is summarized in this paper, we will discuss two of them and review some results of our own, as well as point to other relevant work. These issues are:

- use of multiple types of visual information (cues)
- attention and figure-ground segmentation

Before turning to these topics and to provide a motivation for them we will make some remarks on the general question of how to approach the problem of studying artificial seeing agents.

BMVC 1996 doi:10.5244/C.10.15

2 The Systems Perspective

Studying seeing agents and vision and action from a general perspective is a far-reaching and difficult undertaking. The systems aspects become central, since all components should be regarded in the context of the tasks that the agent performs. Fermüller and Aloimonos propose a model consisting of "procedures for performing visual perceptions, physical actions, learning, and information retrieval, and purposive representations of the perceptual information, along with representations of information acquired over time and stored in memory", topics that cover most of what is being considered in cognitive science, AI and computer vision.

An important question then becomes what kind of approach can be used to investigate such a system in its entirety. Attempts to develop systems that close the loop between perception and action in real environments but in limited contexts certainly exist. One could mention the work on on visual guidance of vehicles which has been quite successful, see e.g. (Masaki, 1995) for an overview. Another example is the work in the ESPRIT-project Vision-as-Process, reported in part in (Crowley and Christensen, 1995), aiming at seeing robots in indoor scenarios. However, none of these projects have considered the problems from the basic perspective that Fermüller and Aloimonos consider. For instance they don't address systems capable of dealing with varying and complex environments and competent to perform a multiplicity of tasks.

In analogy to Brooks (1986), Fermüller and Aloimonos suggest a synthetic approach by which a complex system is built up step by step by adding operational models, thereby including more and more competences. Such a constructive approach is what many researchers in active vision have used, although with less ambitious goals. However, it is far from clear how such a methodology can be scientifically founded.

If the work aims at understanding biological vision, obviously the designed models can be tested against empirical data. If on the other hand, the goal is to design artificial vision systems, there is a need both to devise working systems and to provide some formal analysis of it. Issues on scaling must also be addressed. To date that seems to be beyond the state-of-the-art: systems for which there is an analysis of their functionality, such as the one proposed in (Košecka, 1996), have limited capabilities and address few of the basic issues in vision. Moreover, presented systems have few competences and evidence that they could scale up is missing.

Despite these unanswered questions we believe that a constructive approach is worth pursuing and that it involves design and analysis as well as implementation and experiments. With that in mind we now turn to the two aspects mentioned above.

3 Use of multiple cues

The vision systems we are aiming for in our work should function in complex and unpredictable environments and be capable of performing a repertoire of tasks. Even if particular tasks, such as homing or ground plane obstacle avoidance, may be solved by using one specific type of information, the full system should be able to use whatever information is required and available at the moment, see (Uhlin and Eklundh, 1995; Uhlin, 1996). Hence, questions about cue integration arise.

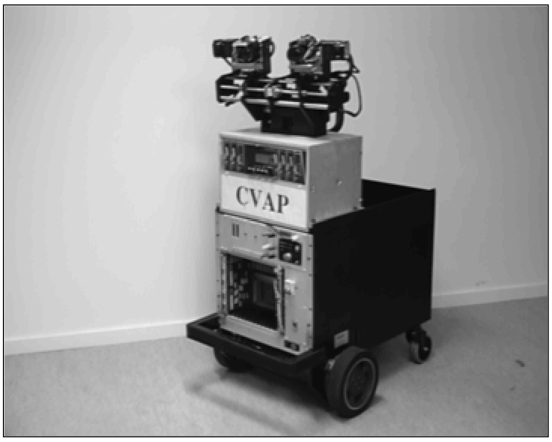


Figure 1: The experimental platform. The KTH head-eye system mounted on a mobile platform.

In human vision it is well-known that the cues often are integrated by averaging. However, from the work of Bülthoff and Mallot (1988; 1991) we know that there also exist mechanisms for overriding and vetoing some information in certain situations.

Machine vision research has mainly been concerned with the "fusion-by-averaging" approach, even though various advanced techniques to introduce more complex behaviors exist, see (Clark and Yuille, 1990) for an extensive treatment. Moreover, the problem has mainly been considered in the context of visual reconstruction.

However, in a purposive and task-oriented approach the *selection* problem becomes important. One particular case is when a cue is or becomes unavailable and the system therefore has to rely on some other cue. An example is when a moving target that is tracked suddenly stops. The motion segmentation no longer works, but stereopsis, or an image or feature based algorithm may still be capable of identifying the target. Similarly, motion and stereo algorithms requiring textures will not work if the observer happens to look at a uniformly colored object. However, in that case the boundary could then be easy to extract.

The knowledge about biological vision can gives us indications for a selection strategy, since it presumably embodies ecological validity. Such an argument suggests that cues from motion, and depth cues from binocular and accommodative disparities in some sense should be given priority. We have applied such approaches in our work on dynamic fixation (see Pahlavan et al., 1996; Uhlin, 1996) and recently in work more directly concerned with the use of multiple cues.

The actual fusion/selection in this work has been done by assuming that consistency over time and between cues is not accidental, and therefore indicates reliable information. By combining this principle with a favoring of motion and three dimensional cues and with computation of uncertainty measures we obtain systems that seem to provide results predicted by our models. Without going into details (which can be found in the references) we illustrate this by two examples, implemented on our mobile platform with the KTH head-eye system mounted on it, Figure 1.

The first, shown in Figures 2-4 demonstrates how the system can rely on motion and binocular disparities as primary cues to keep track of a moving object, but switch to static monocular cues when the object stops.

Each algorithm provides a reliability measure. In Figure 3 we see results of combining stereo-disparity and motion. The masks are computed on a frame-to-frame basis, but the resulting mask shown has been obtained by using a simple hysteresis-update scheme over time. The reliability measure is calculated from the temporal mask-updating algorithm. We favor a compact mask and no sudden area-changes.

By exploiting the reliability measure we can detect that the algorithm is breaking down somewhere between frame 79 and 82.

In Figure 4 we see results using static monocular features more precisely using the size of the dot pattern[1]. This information is not proper for segmenting out the object satisfactory in the beginning of the sequence since the background has the same size of the dot pattern as the object there. Around frame 76 it can be observed that the two algorithms produce masks overlapping very well. From this indication that the second algorithm is producing a reliable result it takes over when the first algorithm breaks down.

In the example in Figure 5 (from Bräutigam et al., 1996), we illustrate coincidence between two types of information. L-junctions indicate planar surface patches. In binocular viewing a matched pair of L-junctions provide one positional disparity (between the intersection points) and two orientation disparities (between the edge elements). In combination with partial information about the viewing geometry, this makes it possible to compute th local surface orientation (Wildes, 1991; Jones and Malik, 1992). This cue is used as a hypothesis for an occurrence of a planar patch which is tested for support by in this case looking at pairs of L-junctions which could stem from right angle bounding a rectangular. In the example, where such features are abundant, the obvious planar surface pops out clearly without reconstruction of the whole scene. We refer to the cited paper for details.

4 Attention and Figure-Ground Segmentation

An agent using vision to guide its actions need to selectively derive information required to solve the tasks at hand. In active vision research it is generally assumed that this selection occurs at an early stage. This leads to the ubiquitous problem of figure-ground segmentation and also to issues on attention.

Let us first make some remarks on attention. In recent years mechanisms for attention has been extensively studied by computer vision researchers, see

[1] This algorithm is only operating in a region of interest area obtained from the former described algorithm.

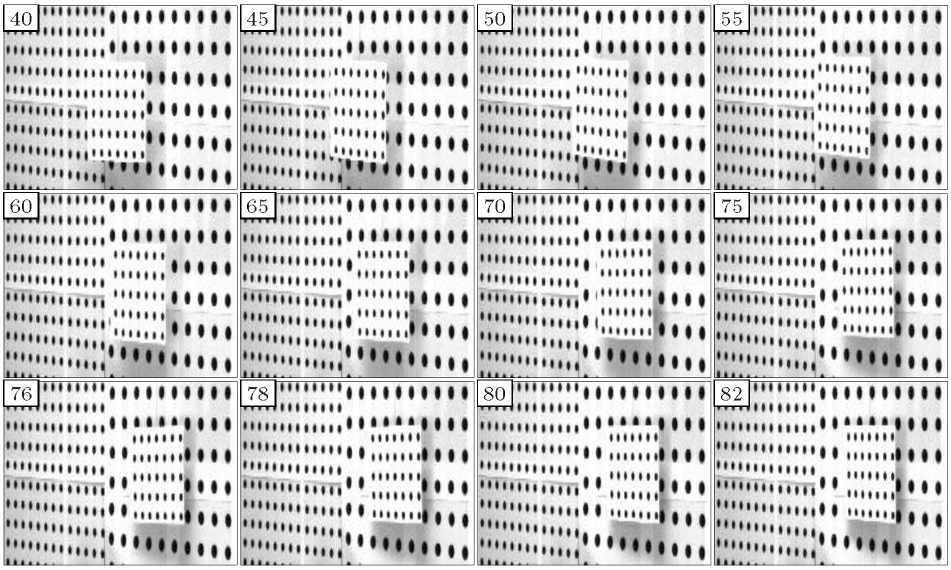


Figure 2: Original sequence, showing a white paper rectangle moving from left to right. The rectangle stops moving in the end of the sequence.

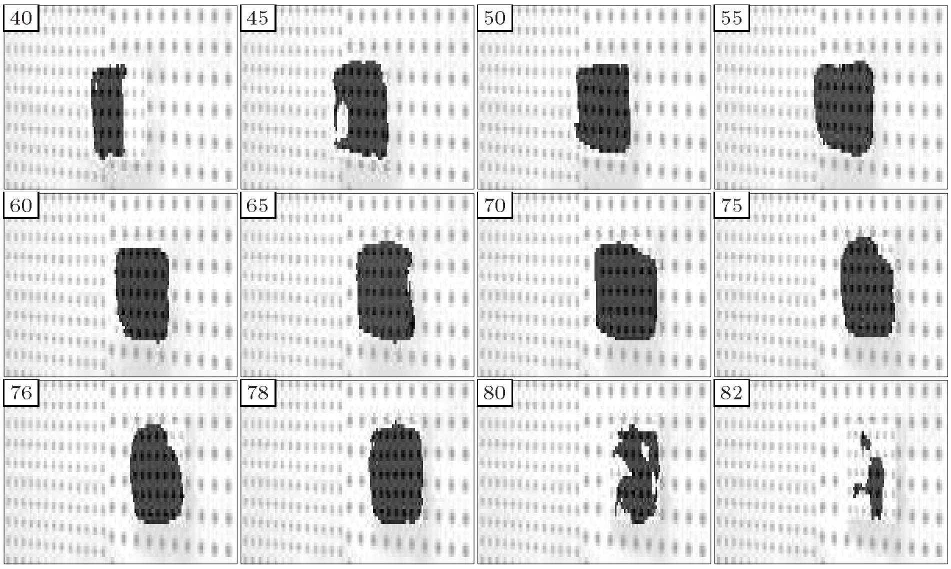


Figure 3: Segmentation using stereo disparity and motion. The algorithm is breaking down somewhere between frame 79 and 82, and the mask changes dramatically.

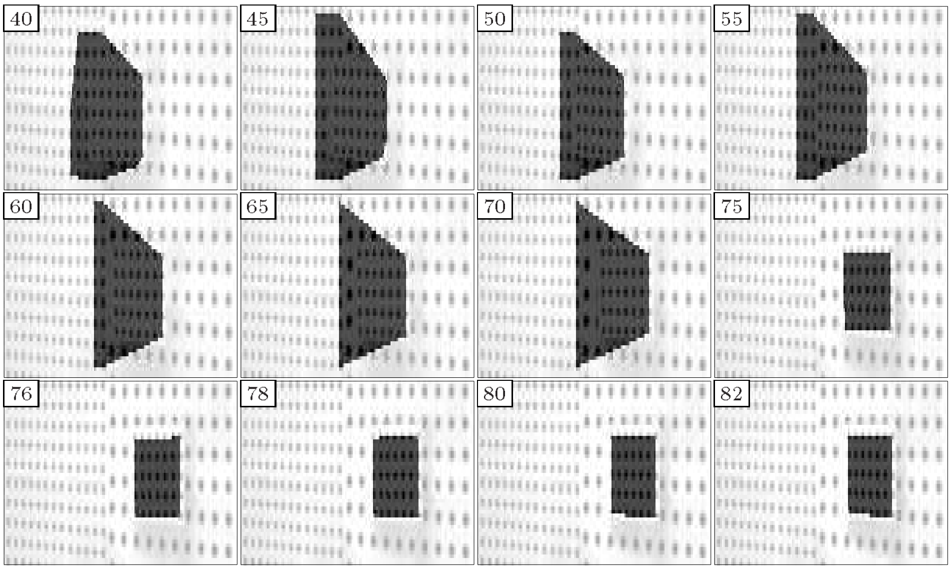
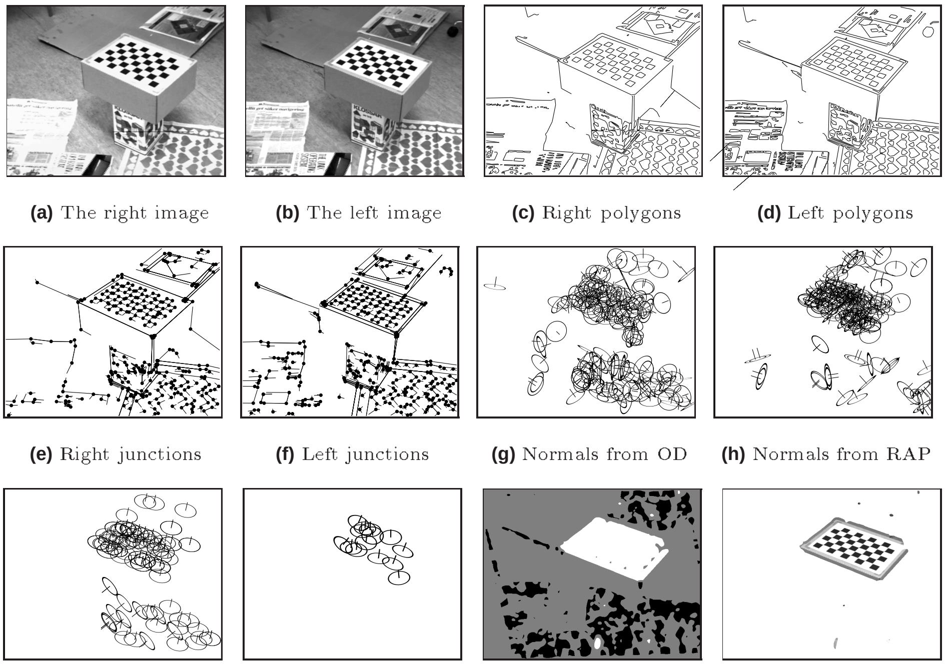


Figure 4: Segmentation using only static monocular features (dot size). Around frame 76-78 the two algorithm produce masks overlapping very well. Hence, when the mask in Figure 3 changes dramatically the system ascribes this mask to the object.

e.g. (Tsotsos et al., 1995; Clark and Ferrier, 1988). There is also a wealth of results on ways to quantify and measure salience. In our work we have built upon these results and particularly emphasized the effectiveness of motion and three-dimensional cues. However, even if mechanisms for attention have been developed, based on e.g. winner-take-all schemes, little is known about how to apply these when an agent is to solve a set of (pre-attentive and attentive) tasks, i.e. the problem of guiding attention in a purposive way. We feel that this problem has to be addressed from the systems perspective, and empirical validation of such models implies substantial implementational work.

In our work (see e.g. Uhlin et al., 1995; Maki et al., 1996) we have demonstrated that three dimensional objects can be detected and segmented out using motion and binocular disparities coupled with estimation and inhibition of background motion caused by egomotion. In fact, we have shown that this can be done by coarse and therefore fast and reliable techniques, although our work so far only applies to textured objects. We have combined these methods with fixation to obtain a system that can detect and hold gaze on an object, inhibiting its background. In principle, this provides time for other algorithms to process the image of a single object, without much disturbance from structure and events in the background.

An example is shown in Figures 6 and 7, from (Maki et al., 1996). The criterion for holding vs shifting gaze is based on habituation: after holding gaze on the same target for n (= 10) frames it is inhibited and a saccade is made. Between competing targets the system generally selects the closest one. These criteria are here hardwired into the system. The aim is to show that this information is indeed



(a) The right image

(b) The left image

(c) Right polygons

(d) Left polygons

(e) Right junctions

(f) Left junctions

(g) Normals from OD

(h) Normals from RAP

(i) Supported clusters

(j) Verified matched normals from the supported cluster

(k) Verified plane (white) and obstacles (black)

(l) Verified plane mask applied to original image

Figure 5: Both local cues, orientation disparity and right angle pairs, give rise to several different surface orientations, as shown in images (g), and (h). But when selecting surface orientation clusters that are supported by both cues, only two clusters remain as shown in image (i). The matched normals from the first cluster that were verified to satisfy a plane projectivity are shown in (j), and the corresponding image mask is shown in (k) and (l).

computable from the cues employed. Note also that the system immediately after a gaze shift picks ups parts of the background, but that these disappear since they are grouped with the background and inhibited. The errors are mainly related to the fact that the ranges of disparity and horizontal flow are fixed.

An example is shown in Figure 6. It includes again three persons in front of a cluttered background: the person A, moving in front from the left to the right throughout the sequence, the person B, standing on the right hand side, and the person C, walking further back from the left hand side. Every 5th frame is shown (images are taken at frame rate 25 Hz). In Figure 7 the target masks are shown. The frames are numbered from 85 to 140. An attentional shift occurs periodically after 10 frames. More specifically:

Frame 85-95: Being the only moving object, A is selected as the target for attention. This period continues even longer than the fixed duration since no other candidate creates sufficient movement to be attended to.

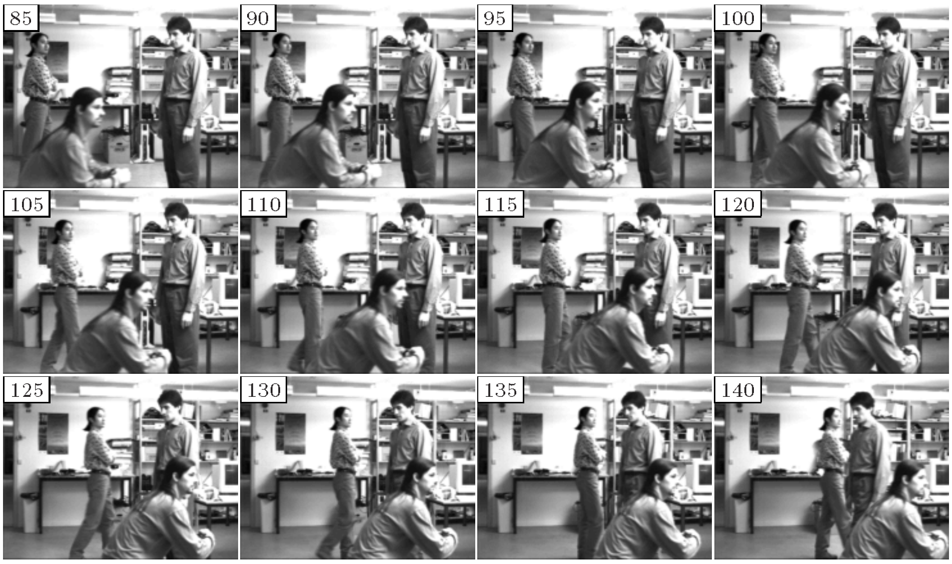


Figure 6: An example sequence with 3 persons taken by a stationary binocular camera head. Top-left to bottom-right. Every 5^{th} frame of the left image is shown (40 ms between frames).

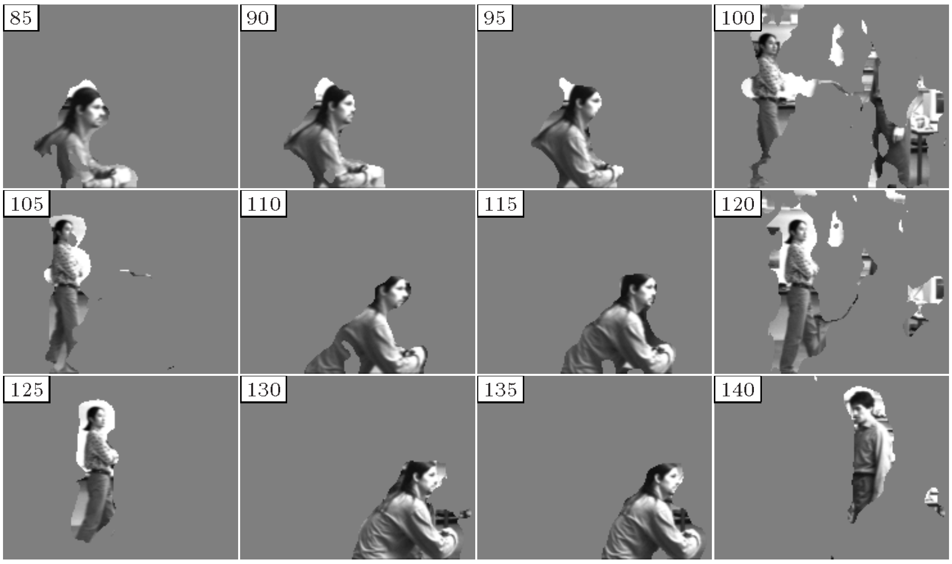


Figure 7: Target masks computed for the sequence in Figure 6. Top-left to bottom-right. The frames are numbered from 85 to 140. Every 5^{th} frame is shown (40 ms between frames).

Frame 100-105: The attention is shifted to C whose movement begins to be recognized. For 10 frames C is kept as the target in pursuit mode.

Frame 110-135: The attention is shifted back and forth after fixed durations. First back to A, then to C and again to A. During this period B with little movement is not among the candidates for attention.

Frame 140: This time attention is shifted instead to B that is finally starting to move in closer to the observer than C. The depth-based criterion applies here.

A particular feature of the approach in the previous example is that the system is capable of first shifting gaze to a moving target and then being able to change to pursuit mode. An example showing this in the case of a static observer (i.e. the platform is static) is given in Figure 8, from (Uhlin, 1996).

It shows, seen from one of the cameras, the system's reaction to a moving target that enters the field of view through a door. Initially the lamp-pole is fixated, but since this is stationary, the attention is switched to the incoming person. The motion in the periphery (left) is detected and a quick gaze-shifting saccade is made to center the person. The saccade utilizes both neck and camera movements to perform the saccade. During the saccade the image is severely blurred due to motion. The saccade is completed in about 3 frames (120 ms). At that time the target becomes reasonably free from motion blur.

As the person enters the door he continues to move across the room. The same figure shows the result of the pursuit movements that follow the initial gaze-shift saccade. The amount of motion may easily be seen from the white lamps that are in the top of the image. Notice how the light condition, the environment, and the motion of the person changes during the pursuit movement. The pursuit is successful in spite of these changes.

The systems and control structure in all these examples is described in detail in the references and will be presented in the talk. The experiments are of a feasibility nature and much remains to be done before more conclusive results can be derived. However, in the spirit of a synthetic approach they provide building blocks that can be used in devising and analyzing more situations and tasks. They can naturally also be applied together with established results on visual recovery.

5 Conclusion

In this paper we have argued that studying vision in the context of action should be done from a systems perspective, supporting ideas elaborated by Fermüller and Aloimonos (1995). We contend that this implies that we need to approach the problem in a constructive manner and actually implement systems, to obtain a way of analyzing the problems arising in real, natural environments. We focused on some issues that we regard as important: use and integration of multiple cues, attention, and figure-ground segmentation. In the paper we have shown examples of a system performing such operations from our own work. The talk will provide a more detailed presentation of these results and their relation to the general issues.

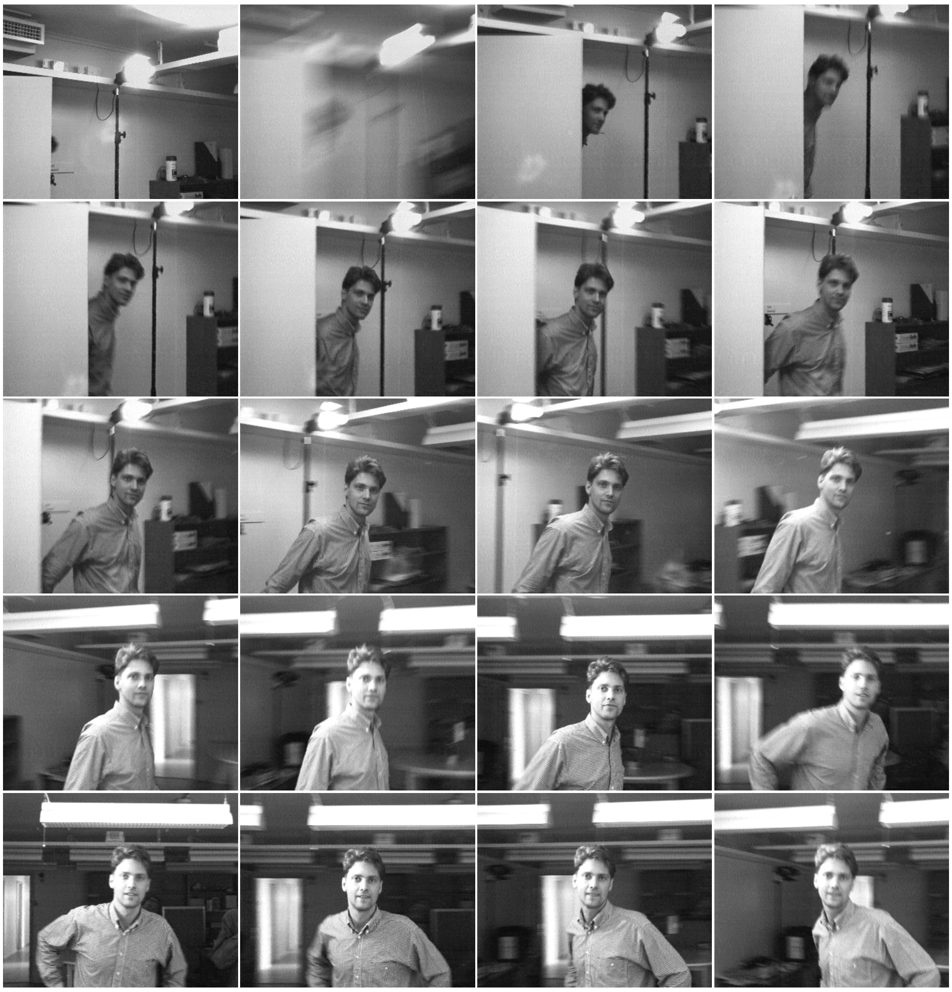


Figure 8: Door watch. A person is entering through a door. An initial saccade catches the person and smooth pursuit is performed as he moves through the room and then turns to move the other way. Every 3^{rd} frame is shown.

References

Aloimonos, Y. (1990). Purposive and qualitative active vision, *Proc. DARPA Image Understanding Workshop*, pp. 816–828.

Aloimonos, Y., Weiss, I. and Bandyopadhyay, A. (1987). Active vision, *Proc. 1st International Conference on Computer Vision*, pp. 35–54.

Bajcsy, R. (1985). Active perception vs. passive perception, *Proc.Third IEEE Workshop on Computer Vision*, IEEE, Bellair, pp. 55–59.

Ballard, D. H. (1991). Animate vision, *J. of Artificial Intelligence* **48**: 57–86.

Bräutigam, C., Gårding, J. and Eklundh, J.-O. (1996). Seeing the obvious, *Technical Report ISRN KTH/NA/P--96/08--SE*, Dept. of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology). Shortened version to appear in Proc. 13th International Conference on Pattern Recognition.

Brooks, R. (1986). A robust layered control system for a mobile robot, *IEEE Journal of Robotics and Automation* **2**: 14–23.

Bülthoff, H. H. (1991). Shape from X: Psychophysics and compution, *in* M. Landy and J. A. Movshon (eds), *Computational Models of Visual Processing*, MIT Press, Cambridge, Ma, pp. 305–330.

Bülthoff, H. H. and Mallot, F. A. (1988). Integration of depth modules: stereo and shading, *Journal of the Optical Society of America A* **5**: 1749–1758.

Clark, J. J. and Ferrier, N. J. (1988). Modal control of an attentive vision system, *Proc. 2nd International Conference on Computer Vision*, pp. 514–523.

Clark, J. J. and Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems*, Kluwer Academic Publishers, Boston, Ma.

Crowley, J. L. and Christensen, H. I. (eds) (1995). *Vision as Process*, Basic Research Series, Springer Verlag, Berlin.

Eklundh, J.-O. (1995). Trends in active vision, *in* J. van Leuween (ed.), *Computer Science Today*, Vol. 1000 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 515–517.

Fermüller, C. and Aloimonos, Y. (1995). Vision and action, *Image and Vision Computing* **13**: 725–744.

Jones, D. G. and Malik, J. (1992). Determining three-dimensional shape from orientational and spatial frequency disparities, *Proc. 2nd European Conference on Computer Vision*, Vol. 588 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Santa Margherita Ligure, Italy, pp. 661–669.

Košecka, J. (1996). *A Framework for Modeling and Verifying Visually Guided Agents, Analysis and Experiments*, Ph. D. dissertation, Dept of Computer and Information Science, Univ of Pennsylvania.

Maki, A., Eklundh, J.-O. and Nordlund, P. (1996). A computational model of depth-based attention, *Technical Report ISRN KTH/NA/P--96/05--SE*, Dept. of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology). Shortened version to appear in Proc. 13th International Conference on Pattern Recognition.

Masaki, I. (ed.) (1995). *Proc. Int. Symp. on Intelligent Vehicles '95*, Detroit, Mi.

Nelson, R. C. (1991). Vision as intelligent behavior – an introduction to machine vision research at the University of Rochester, *International Journal of Computer Vision* **7**: 5–10.

Pahlavan, K., Uhlin, T. and Eklundh, J.-O. (1996). Dynamic fixation and active perception, *International Journal of Computer Vision* **17**: 113–135.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Y. Lai, N. D. and Nuflo, F. (1995). Modeling visual attention via selective tuning, *J. of Artificial Intelligence* **78**: 507–545.

Uhlin, T. (1996). *Fixation and Seeing Systems*, Ph. D. dissertation, Dept. of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology).

Uhlin, T. and Eklundh, J.-O. (1995). Animate vision in a rich environment, *Proc. 14th Int. Joint Conf. Artificial Intelligence*, Montreal, Canada, pp. 27–33.

Uhlin, T., Nordlund, P., Maki, A. and Eklundh, J.-O. (1995). Towards an active visual observer, *Technical Report ISRN KTH/NA/P--95/08--SE*, Dept. of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology). Shortened version in Proc. 5th International Conference on Computer Vision pp 679–686.

Wildes, R. P. (1991). Direct recovery of three-dimensional scene geometry from binocular stereo disparity, *IEEE Trans. Pattern Analysis and Machine Intell.* **13**: 761–774.