Developing an Active Observer

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

October 23, 1995

Abstract

Seeing robots are usually aimed at functioning in real environments. Hence the figure-ground problem is essential. We argue that for a seeing robot, capable of actively fixating and holding gaze on objects in three dimensions, this problem is manageable, in particular if multiple cues can be used. An important point here is that a seeing robot should be able to utilize what the environment offers, rather than relying on a predetermined set of features.

Furthermore, if there are early processes for figure-ground segmentation the local statistics of an observed (yet unknown) objects should be simpler than that of the entire scene. That suggest that processing to derive further properties, of e.g the shape or motion of that object, or even recognizing, could be based on other features than those used to segment out the object. Such a view fits well with recent theories about view-based recognition. It also allows efficient implementations in terms of a visual-front-end, since both the target selection and the target analysis can be based on the output from such a layer, even though different features are used.

1 Introduction

Vision as we know it from biology is an active process. The seeing agent uses vision as well as its other senses to guide its behaviors. Current research in active computer vision studies a range of such systems, such as the prey catching and threat avoiding system of Arbib and Liaw [1] that uses very simple perception, and systems such as the one presented by Kosecka *et al.* [4], that involves rather complex perceptual processes.

A major emphasis in active vision research has not surprisingly been on problems similar to those traditionally studied in computer vision. Typical efforts deal with navigation, manipulation and recognition tasks in indoor and outdoor environments usually encountered by man. Such problems have for a long time been addressed without reference to vision as an active process. However, despite considerable progress on computational techniques for deriving information from the scenes from images of it, and despite striking successes in cases such as vision-based vehicle guidance, one can hardly claim that "seeing robots" capable of advanced and flexible behavior in rich environments exist. This is especially obvious when human performance is used as the (implicit) yardstick for measuring success, which is often the case.

One particular and ubiquitous problem that in the authors' view has prevented the application of state-of-the-art and theoretically well-founded techniques of traditional computer vision to real-life situations is what is generally called the figure-ground problem. A major reason for this is that each such technique makes very precise assumptions of what features are needed, while consistent extraction of these features often turn out to be very difficult in complex environments. In fact, they may not even be available. On the other hand we know that humans most of the time experience few problems in segmenting out objects in the surrounding world. Of course, these objects are of many types and sizes, some complex and composite, while others are simple and only parts of larger entities. As we know, what constitutes an object also depends on the task at hand and the context. Nevertheless, humans have few problems with this, while bootstrapping computer vision algorithms for the analysis of the same type of scenes seems extremely difficult.

In this paper we argue that such problems have viable solutions in active vision, which assumes a system that acquires visual information by looking at a *scene* rather than at images of it - it goes beyond pictorial vision. We suggest that this leads to an approach considering the figure-ground problem in the three dimensional world, and separating it from the later stages of the analysis. Well-known techniques are still used, but by adapting to what the world offers we will in fact facilitate their applicability by limiting the context. Our arguments are based on several observations that we will elaborate on in subsequent sections. We will then present some experimental evidence supporting our view and discuss

*The authors want to thank Kourosh Pahlavan for valuable discussions on the topic of the article. The reported work has been supported by a grant from the Swedish National Board for Technical and Industrial Development. This support is gratefully acknowledged.

their consequences.

1.1 The active observer and the figure-ground problem

Consider a person moving around in the world, while looking at various locations and things in his way. Objects will then due to his own or their motion enter and leave his field of view. He will sometimes shift his gaze to such objects to find out, e.g. if their trajectory crosses his path, or, to determine what kind of things they are. In fact, he may be looking for specific types or instances of objects, for example they may be obstacles to avoid, or something he needs.

Certain capabilities that this observer displays are of importance for a seeing robot as well. First, the observer is capable of singling out things that stand out from their immediate surroundings. Secondly, he can hold these things in his field of view long enough to identify and derive various properties of them.

The first problem is the mentioned figure-ground problem, but it can be noted that often something is considered as an object just because it stands out as a three-dimensional entity. Hence, cues indicating that are essential. This is not surprising, since camouflage in principle is impossible in three dimensions. It is well-known that motion, binocular disparities, and blur give such information with greater certainty than monocular image cues. Hence, early use of three dimensional information seems important, as is also suggested in the study of human attention by Nakayama and his co-workers, see e.g Nakayama and Silverman [6] and Shimojo *et al.* [9].

In active vision such information is readily available. Binocular head-eye systems capable of fixation can, as was pointed out early by Ballard [2], observe three-dimensional structure in the world in relation to the fixation point. By shifting fixation this can be applied to various locations in the scene. That effect is present both when binocular disparities and motion parallax is considered. As discussed in Pahlavan *et al.* [8] blur as a depth cue is conveniently available to an active system capable of controlling its optical parameters.

When a target is detected, an active system can hold gaze on it. This can be done by image tracking, but pursuit movements are needed if the target tends to move out of the field of view. There are several important consequences of such a capability. First, the earlier attention step has created a region of interest that corresponds to an actual physical target (if that step has accomplished its task). The characteristic features of that target can hence be found in this region. It is reasonable to assume that the statistics of various features are simpler there than over the scene as a whole. Furthermore, if the target is masked out, then any typical feature of the object can be used to characterize or recognize it, whether this feature is common in the background or not. This becomes particularly important in a dynamic situation, when the background will vary independently of the object. Since the target is seen for some time, one can also assume that certain characteristics can be ascribed to it. Note that the target has an identity before one knows what it is. These characteristics can then be used to compute and maintain information about e.g. its shape and motion, but also to establish that it is the same object when it stops or changes motion pattern, or when only a fraction of it is seen due to occlusion: The target that's singled out forms its own model. Moreover, what is used to analyze it further, is in principle independent of what was used to identify it as being an object.

To effectively apply a technique like the one we've outlined a third required capability becomes obvious: Many different cues must be computed concurrently, and the system must be able to choose those that at a particular instance of time are available and distinctive. This suggests a *systems approach* to the problem, in which many different processes are integrated and controlled over time. It also suggests that computations made on the image data are reused whenever possible, for efficiency. Recent research on Visual-Front-End models for early vision, see e.g. Gårding and Lindeberg [3], provides possible model for that.

In this paper we'll present experiments that show that this is a promising computational framework in which active vision plays an important role. In principle, the information needed to implement a scheme like this can be derived by traditional methods. Here we will try to show that active vision is especially suited for it.

2 The System Description

The principles described above have been implemented in a system embodying an autonomous mobile observer. This observer should while moving be able to detect and smoothly pursue moving or stationary targets binocularly while maintaining vergence. Currently, motion and binocular disparities are used in the figure-ground separation step, but the architecture is open for additional cues. So far fundamental skills in terms of fixation, target pursuit and target discrimination, have been implemented. The basic system partly runs in real-time on an existing mobile platform, and partly as a post-processing stage working on images captured in real-time during the execution of the former processes.

Mechanisms for discovering moving targets provide means of shifting attention. In other words, our system has two components, one to maintain attention, and another one to find and select new locations to attend to. Moreover, these function in parallel, because the system would otherwise be without choices. They are implemented in the form of a pursuit and a motion detection mechanism, thus obtaining what we believe is the most basic behavioral level for an active observer (Figure 1). (Currently, motion provides the only cue to changing attention, but other cues can easily be incorporated.)

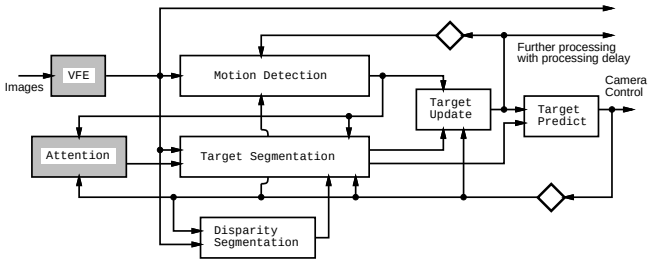

Figure 1: The system implementation is shown schematically. (The diamonds indicate a one frame delay.)

A note on terminology. The term *target* will denote the portion in the image which is currently attended and supposed to represent a scene-object. A *target model* is used, which is coded as an image mask, including pixels that are considered as belonging to the target. All segmentation modules produce similar masks after a thresholding step. These masks are used for the integration.

2.1 Attention and Smooth Pursuit

A key feature of the system lies in its ability to smoothly pursue an arbitrary target that is described by its location, extent and visual appearance. No a priori knowledge about the target's visual appearance, i.e. texture or shape, is built into the system, although additional constraints or knowledge can be included dynamically as information becomes available, and similarly that such information can be excluded as it becomes obsolete. The features that are seen, are assumed by the pursuit module to change smoothly. Non smooth changes will trigger attentional mechanisms.

The implementation is based on a coarse to fine correlation scheme similar to the one reported in Pahlavan *et al.* [8]. This technique works well when there are no occlusions. We will introduce an extension handling occlusions as well.

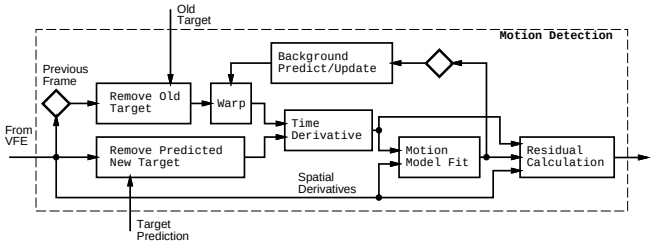

Figure 2: The motion detection is shown schematically. This is a detailed description of the `Motion Detection` box in Figure 1. (The diamonds indicate a one frame delay.)

To handle occluding objects and distracting things in the background, motion detection is integrated to filter out parts of the scene that can be parts of the target, namely those that are moving. This is provided in the system by a connection from `Motion Detection` to `Target Segmentation`, see Figure 1. We will see later in the experiments that this is not enough in many cases when there are occluding objects that are themselves moving, since they will also be detected as moving. Binocular disparity can in these situations aid in depth discrimination by also providing a clue to where occluding objects may lie.

2.1.1 Motion Detection

Motion detection allows the system to shift its attention to another location, if the system "looses interest" or fails to pursue the current target. Control mechanisms of the former type are not present in the current system, but could easily be included. Motion detection functions continuously during pursuit and is used in selecting points where the target is lying (Figure 2).

Using the brightness constancy constraint a global affine image velocity model is fit to the gradient and time derivative information (see `Motion Model Fit` in Figure 2). Two steps involving feedback are included to account for object motion and large background motion.

- The predicted and previous position and extent of the target are used to mask out parts of the image which likely belong to the object, so that they do not affect the calculation of the affine parameters for the background. See feedback into `Motion Detection` in Figures 1 and 2. This is a crucial step to avoid iterations that otherwise would have been necessary to get reasonable performance even if targets are large. We could say that the iterations are performed in time instead of in a single frame.

- The accumulated affine parameters are used over time to cancel out the majority of the time difference, see feedback into `Warp` in Figure 2. Thereby we can avoid iterations in fitting the affine model, and instead iteratively refine our motion model in time. This is increases the computational speed and decreases the subsequent lag of the pursuit mechanism. We have here an example of how one can tradeoff the performance of a single algorithm for better overall performance.

Once the fit is made to the background, and the affine parameters are available, the residual normal image velocity is calculated in the entire image, `Residual Calculation` in Figure 2. The obtained residual is thresholded and high residuals are considered as possible independently moving areas in the scene. The threshold on the residual is adaptive and is set relative to the difference between target and background motion.

2.1.2 Target Segmentation

The aim of target segmentation is to determine which parts of the scene are moving consistently with what is currently believed to be the target. This is the typical situation in connection with occlusions, when the target gradually disappears, to later emerge on the other side of the occlusion. It also supplies the system with the ability to discriminate between several moving targets, found by the motion detection, and distinguish parts

of the scene that lie at the same depth, found by the disparity segmentation (Figure 3).

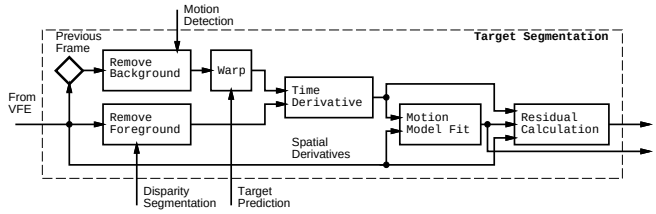

Figure 3: The target segmentation is shown schematically. This is a description of the `Target Segmentation` box in Figure 1. (The diamonds indicate a one frame delay.)

The calculations on the target are performed in analogy with what is done for the background motion, with the advantage that most of the computations spent on the background can be used also here (spatial derivatives and smoothed images from the VFE). By sharing these data we greatly increase the performance.

Analogous to the motion detection in Section 2.1.1, the residual normal image velocity is calculated and an adaptive thresholding is applied.

2.1.3 Disparity Segmentation

The system in addition computes relative depth information. Having a disparity measure is also necessary to obtain vergence.

The disparity computation is carried out with a phase based method which fits well into the VFE framework since the same data can be used for vergence and pursuit movements. It also performs well in a real-time implementation for vergence on the KTH-Head-Eye system, see Uhlin *et al.* [10].

The object of disparity selection is to select the disparities that belong to the target in the presence of disparities that arise from other locations in the scene. The method is described in Maki *et al.* [5] and has been further developed to take advantage of the feedback obtained when the system is run in a closed loop together with other cues to follow a single target, to ensure that version and vergence components in smooth pursuit remain consistent. The basis for selection is the predicted location and extent of the target.

2.2 Integration

The integration is performed mainly in `Target Update` (Figure 1), but in the `Target Segmentation` (Figures 1,3), motion detection information and disparities

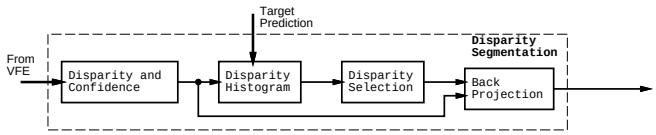

Figure 4: The disparity segmentation is shown schematically. This is a detailed description of the `Disparity Segmentation` box in Figure 1.

are also integrated. The target extent is kept in a mask. The integration produces an updated target mask using masks produced in the paragraphs "Motion Detection" and "Target Segmentation" in Section 2.1.1. There are two parts needed to update the target mask, one to exclude pixels and and one to include pixels.

Target areas that do not get support from either motion detection or target segmentation are excluded from the target model. Also the disparity module detects areas in the scene that lie in front of the pursued target, which are then excluded from the target model.

Areas in the scene that are detected as both moving independently, detected by motion detection, and are moving consistently with the target image velocity model from the target segmentation, are added to the target model.

3 Real-Time Implementation

Presently we have implemented parts of what has been described above to perform real-time visual processing coupled to motor control of the KTH Head-Eye system. Hence, there exist mechanisms for pursuit, motion detection and disparity estimation running at 25 Hz. See Uhlin *et al.* [10] for more details.

4 Experiments and Results

Figure 5: Resulting target pixels as extracted by the system (after integration and update) from the sequence shown in the top row. Using motion detection and an affine image velocity model for the target we can find areas that belong to the target such that occluding objects are removed and a correct pursuit is performed although we have partial occlusion. The pursued target is marked by the small rectangle. The small rectangle marks a fix point on the background as computed by the optokinetic mechanism. The pursuit is performed at 25 Hz, and shown are every 60^{th} frame.

We will briefly indicate how the system performs in a few different situations, both in the presence of all cues and when some cues are not present. The target in most examples undergo complex movement, including rotations and movements towards and from the camera, while the observer undergoes ego-motion. No knowledge about the motion is assumed. Additional examples are presented in Uhlin *et al.* [10], where also more details are given.

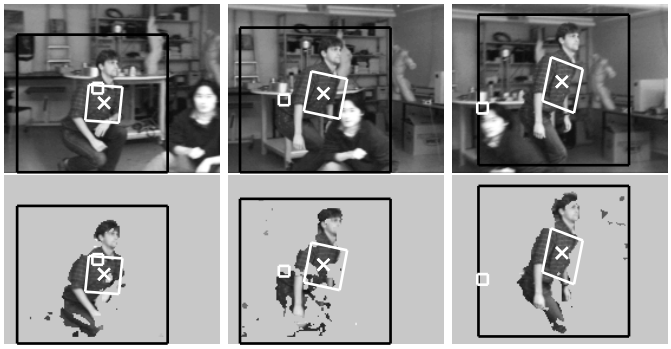

Figure 6: A sequence captured during pursuit. The pursuit is performed at 25 Hz, and shown are every 18^{th} frame.

Figure 7: Target pixels as extracted by the system from the sequence shown in Figure 6, but without the disparity cue. The attention shifts to the other moving person. Every 18^{th} frame is shown here.

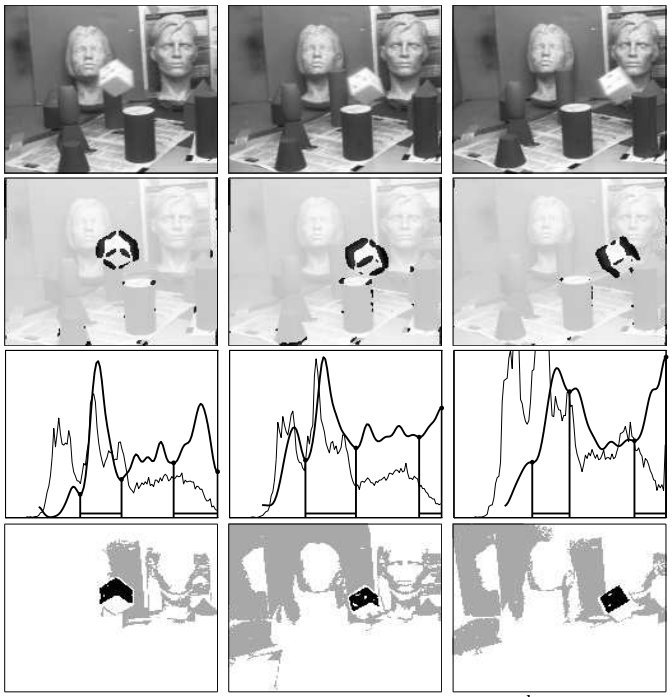

Figure 8: 1^{st} row shows original images. 2^{nd} row shows motion mask obtained with simple motion-algorithm. 3^{rd} row shows grayscale histogram. The histogram for the original image is indicated with a thin line. The histogram for just the area in the motion mask is indicated with a thick line. Straight lines indicate the ranges used to threshold out the object, respectively the background. Compare the significant difference between the two histograms. The 4^{th} row shows the thresholding results. Object marked in black, background marked in gray. Only the connected regions overlapping the motion mask most in respective thresholding range are shown.

The first example, Figure 5, shows a laterally moving target, a human, the derived masks, a fixed point in the background and the estimated local motion. That example contains only one moving target, while Figures 6 and 7 demonstrate the multiple target case. By using disparities together with motion the system can maintain the identity of the observer target also when other targets appear. Figure 8, finally, shows how properties of the target, here simple luminance features, indeed can be determined when it is singled out.

5 Discussion

We have argued that an active observer by dynamically fixating in the three dimensional world can single out objects of interest. This suggests an approach in which the figure-ground problem and subsequent processing in principle is independent, although based on similar information. It also suggests that known methods for e.g. feature extraction and recognition may be facilitated, since local statistics is simpler.

References

[1] Arbib, M.A. and Liaw J-S. (1995). Sensorimotor Transformations in the World of Frogs and Robots. Journal of Artificial Intelligence. **72**, 53–79.

[2] Ballard, D.H. (1991). Animate vision. Journal of Artificial Intelligence. **48**, 57–86.

[3] Gårding, J. and Lindeberg, T. (1995). Direct Computation of Shape cues Based on Scale-Adapted Spatial Derivative Operators. International Journal of Computer Vision. To appear.

[4] Kosecka, J, Christensen, H. I., and Bajcsy, R. (1995). Discrete Event Modelling of Visually Guided Behaviors. International Journal of Computer Vision, **14**, 179–191.

[5] Maki, A., Uhlin, T., and Eklundh, J.-O. (1994). Disparity selection in binocular pursuit. In *Proc. 4th IAPR Workshop on Machine Vision Applications*, pp. 182–185.

[6] Nakayama, K. and Silverman, G. H. (1988). Serial and parallel processing of visual feature conjunctions. *Nature*, no. 320, 264–265.

[7] Pahlavan, K. Uhlin, T. and Eklundh, J-O. (1993). Active vision as Methodology. Active Perception. Aloimonos, Y. (ed.): Lawrence Erlbaum Associates, Hillsdale, NJ.

[8] Pahlavan, K., Uhlin, T., and Eklundh, J.-O. (1993). Dynamic fixation. In *Proc. 4th ICCV*, pp. 412–419, Berlin, Germany. IEEE Computer Society Press.

[9] Shimojo, S., Silverman, G. H., and Nakayama, K. (1988). An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature*, no. 222, 265–268.

[10] Uhlin T., Nordlund P., Maki A., and Eklundh J.-O. (1995). Towards an active visual observer. Tech. Rep. 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.