A Computational Model of Depth-Based Attention

Atsuto Maki,* Peter Nordlund and Jan-Olof Eklundh

Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computing Science
Royal Institute of Technology, S–100 44 Stockholm, Sweden
Email:maki@bion.kth.se, petern@bion.kth.se and joe@bion.kth.se

Abstract

We present a computational model for attention. It consists of an early parallel stage with preattentive cues followed by a later serial stage, where the cues are integrated. We base the model on disparity, image flow and motion. As one of the several possibilities we choose a depth-based criterion to integrate these cues, in such a way that the attention is maintained to the closest moving object. We demonstrate the technique by experiments in which a moving observer selectively masks out different moving objects in real scenes.

Keywords: *Attention, Relative depth, Cue integration, Pursuit, Saccade, Target selection*

1 Introduction

The notion of attention plays an important role in biological vision. In recent years, and especially with the emerging interest in active vision, computer vision researchers have been increasingly concerned with attentional mechanisms as well, see [9, 11, 7] for a few examples. The basic principles behind these efforts are greatly influenced by psychophysical research. That is the case also in the work presented here, which adapts to the model of Treisman [8], with an early parallel stage with preattentive cues followed by a later serial stage where the cues are integrated.

In our case, we base the model on disparity, image flow, and motion. A schematic diagram of our framework is shown in Figure 1. In this scheme cue integration and attention over time are essential aspects. Part of the cue integration work has appeared in [10]. The contribution here is that we show that the system can attend to different targets in a purposive way in a cluttered environment. A second key point in this context is the use of depth information, as suggested is done in human vision by Nakayama and Silverman [4]. The computation of precise depth information is generally a time consuming task. A third important point of this work is therefore that a functioning system capable of selectively attending different objects can be obtained with rather simple algorithms allowing fast implementations. This is demonstrated by experiments in which a moving or stationary binocular observer (a mobile platform with a head-eye system) selectively masks out different moving objects in real scenes and holds gaze on them over some frames. The selection criteria are here based on nearness and motion, but could in our open architecture be of any type.

*Maki's work was performed while a Ph. D. Student at CVAP, Royal Inst. of Technology. Present address: TOSHIBA Kansai Research Laboratory 6-26, Motoyama-Minami-cho, 8-Chome, Higashinada-ku, Kobe, 658, Japan.

2 Early modules

This section describes the preattentive cues employed in the early parallel stage: stereo disparity, image flow and motion detection, which are integrated in the later serial stage.

2.1 Stereo disparity

Relative depth, that plays a central role in our model, is derived from a dense disparity map. As disparity estimator we employ a phase-based algorithm which has the advantages of low computational cost, stability against varying lighting condition and especially of allowing good direct localization of the estimated disparity. The disparity estimation algorithm employed here is found in [1]. A target mask is produced by back projection of a selected target disparity and the process of disparity selection is based on histogramming and disparity prediction. The idea is to slice up the scene according to relative-depth and then segment out the part of the input image corresponding to the selected target as a mask. See [2] for details of the procedure producing the target mask. A point to be noted is that the resulting

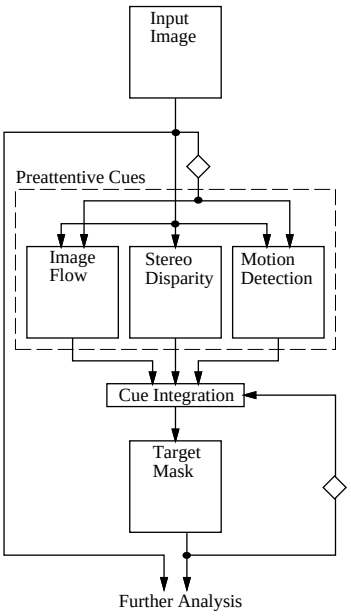


Figure 1. A schematic diagram of the proposed attentional framework. It follows the general concept of visual attention, i.e. the early parallel stage with preattentive cues and the later serial stage where the cues are integrated. The diamonds indicate a one frame delay.

mask may well involve multiple targets if they are observed to be close to each other in depth. Further segmentation among such targets is beyond the performance of the depth cue alone and some additional information sources would be necessary to handle such a situation.

2.2 Image flow

By applying the stereo algorithm to consecutive image frames instead of to a stereo image pair, information of horizontal image flow can be obtained. The image flow cue provides another target mask independent of the depth cue, and those cues are combined in order to deal with complex scenes where multiple target candidates are observed. Information about image flow could be made available in more specific form and as a matter of fact it could be by itself a central cue in terms of attending to moving objects [3]. In our scheme, however, the use of image flow cue is only in one-dimension along the horizontal direction, because by doing so this early module can share identical input with the depth module. This additional module is in our experiments shown to stabilize the attentional performance a great deal, in spite of its simplicity.

2.3 Motion detection

As the third module in the early preattentive stage, a technique for motion detection is employed. The fundamental concept is outlined here. The idea is to exploit the brightness constancy constraint in conjunction with an affine transformation between two consecutive images. Assuming the moving target to be relatively small compared to the background, we compute an affine fit between two consecutive images by posing a weighted least squares minimization problem. Given that the background contains small variations in depth and is far away enough, relative to the motion, the background cancels in the residual image and moving objects appear. A full description of the technique is found in [5].

3 Cue integration

Given information from the early stage in the form of stereo disparity, image flow estimation and detected motion, the role of the later stage is to guide the attention to an appropriate part of the input image. This guidance is achieved by combination of the different early cues in two independent modes, namely the pursuit and saccade modes, each of which produces a target mask. As a criterion to choose the final attentional target mask, depth-based target selection is considered.

3.1 Pursuit mode

The objective in the pursuit mode is to keep attending to the current target and mask the corresponding part of the input image accordingly. The framework of the process in this mode is schematically depicted in Figure 2 using the following notation at frame number k, $T_p(k)$: Target pursuit mask, $T_d(k)$: Target mask based on stereo disparity, $T_f(k)$: Target mask based on image flow, and $T(k-1)$: Target mask in the previous frame. Taking the disparity and flow maps as inputs from the early stage, it returns a target pursuit mask $T_p(k)$ as output.

As described earlier, the disparity target mask $T_d(k)$ is produced by a disparity selection technique based on histogramming and back projection. The information of image flow is processed in an analogous framework, that is, a one-dimensional histogram is constructed for the horizontal flow map and a flow target mask $T_f(k)$ is produced by back projection of a flow parameter that is also selected based on prediction. To summarize, from each of the disparity and flow maps a target mask is produced and those masks are fused with a logical *and* operation into the target pursuit mask $T_p(k)$ so that just the part which is consistent with

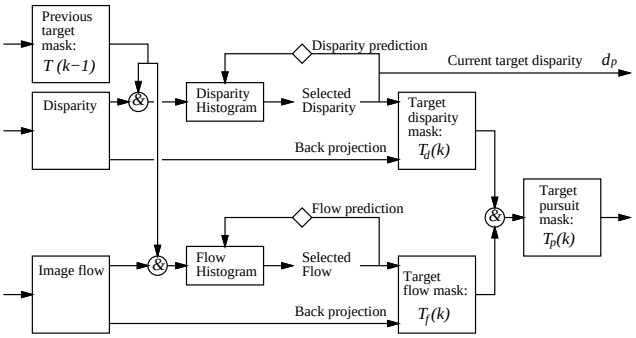


Figure 2. Schematic fbw diagram of the attentional pursuit. It composes part of the "cue integration" in the framework shown in Figure 1. The diamonds indicate a one frame delay in the feedback. The circles with & indicate a logical and operation.

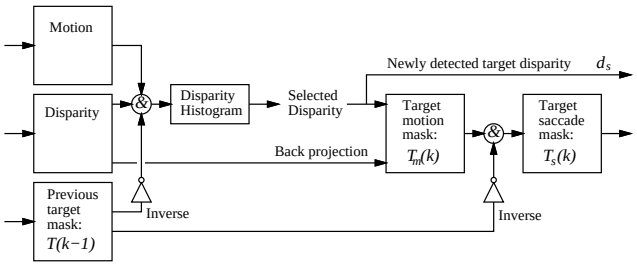


Figure 3. Schematic fbw diagram of the attentional saccade. It composes part of the "cue integration" in the framework shown in Figure 1. The circles with & indicate a logical and operation.

both disparity and flow remains. The process in frame *k* can be formulated as:

$$T_p(k) \quad = \quad T_d(k) \cap T_f(k) \tag{1}$$

3.2 Saccade mode

The saccade mode on the other hand is aimed at disengaging the attention from the current target and shift it to a new one. The framework of the process in this mode is schematically depicted in Figure 3 using the following notation at frame k, $T_s(k)$: Target saccade mask, $T_m(k)$: Target mask based on detected motion, and $T(k-1)$: Target mask in the previous frame. While the disparity cue again plays the central role here, the important feature is that a shift is triggered when a new interesting part in the input is detected. The definition of "interesting part" can be task dependent and any distractor among available alternatives could in principle trigger an attentional shift. Here we have chosen only motion relative to the background, since it provides a strong saccadic cue and therefore allows us to demonstrate our framework.

As is the case in the pursuit mode, a target saccade mask $T_s(k)$ is produced basically by the disparity selection and serial back projection. The previous target mask $T(k-1)$ is, however, utilized differently, i.e. in the saccade mode $T(k-1)$ is inversely applied, so that the current target is inhibited instead of accepted as contribution to the disparity histogram. Besides, the use of disparity information as input is restricted to the part where relative motion to the background is detected. The disparity histogram then carries information just about a newly detected moving target. The process is completed by inhibiting the produced target mask again by $T(k-1)$ to make sure that the resulting target saccade mask $T_s(k)$ does not overlap the former target. The process in frame *k* can be summarized as:

$$T_s(k) \quad = \quad T_m(k) \cap \overline{T(k-1)} \tag{2}$$

It should be noted that the framework of the saccade mode without feedback of the former target mask exactly provides a mode to initiate the process by finding the moving target to attend to. This also applies to picking up a new moving target and to restart the process in case the pursuit mode for some reason lost track of the target, e.g. when the target disappears from the scene.

3.3 Depth-based target selection

The cue integration process described so far provides a pursuit mask $T_p(k)$ and a saccade mask $T_s(k)$ and the choice among those masks is the remaining issue, which is rather task dependent. Some criterion is needed to decide when the saccade should happen or pursuit should continue and thereby to determine the final target mask $T(k)$ in each frame (see Figure 4). depth-based criterion and time transient criterion. While the framework introduced is open to accept different criteria, we here have considered a depth-based attentional scheme where the target that is closer in depth is selected with higher priority, see equation (3). Such a criterion is reasonable for instance for a moving observer that wants to avoid obstacles. The target saccade mask is selected when the newly detected target turns out to be closer, or the current target disappears from the scene, and thus the closest moving object is kept on attended as a target over time.

$$T(k) \quad = \quad \begin{cases} T_s(k), & \text{for } d_s < d_p \\ T_p(k), & \text{otherwise} \end{cases} \tag{3}$$

d_s : Disparity of newly detected target

d_p : Disparity of the current target

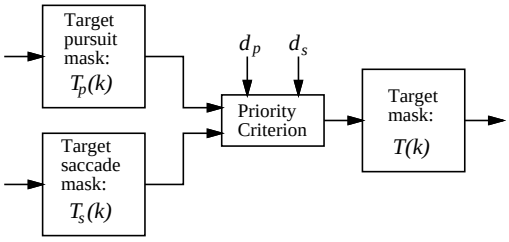


Figure 4. Attentional target. In each frame either target pursuit or target saccade mask is selected as the final target mask according to a criterion.

Figure 5. An example sequence with 3 moving persons taken by a stationary binocular camera head. Every 10th frame of the left image is shown (40 msec between frames).

4 Experiments

Figure 5 shows a sample image sequence seen by a stationary binocular camera head. It includes three persons walking around in a laboratory. Every 10th frame is shown (images are taken at framerate 25 Hz). Disparity maps, horizontal flow maps and detected motion maps are shown in Figure 6 - Figure 8[1]. From both of the relative depth and flow information parts of the maps corresponding to the persons in the scene are recognized in each frame. As the third preattentive cue, information of detected motion, shown in Figure 8, is incorporated in such a way that the closest moving object is kept on attended as the target in each frame, (see Figure 9).

Figure 10 shows another sample image sequence, this time by rotating cameras [6]. It includes two persons walking in a laboratory, one tracked at the center of the image, and the other appearing on the right hand side and passing by in front, while the observing camera head is rotating laterally. Every 10th frame is shown (images are taken at framerate 25 Hz). Figure 11 shows that the motion detection process functions even for a sequence with rotating cameras. The resulting target masks are shown in Figure 12, where it is observed that the closest moving object is kept attended to.

Figure 13 exemplifies the process by which the target mask is selected. Illustrated are (a) the masks restricting the input to the disparity histogram, (b),(c) two histograms in pursuit and saccade modes, and (d) the target mask superimposed on the original input image. They are shown for three consecutive frames, $k-1, k$ and $k+1$ (left, middle and right) to clarify the information flow between frames. A detailed description of the process along the frames is as follows.

[1]The gray scale in those maps is not necessarily consistent throughout the frames since it is scaled in the range between the highest and lowest values in each frame respectively.

Frame $k-1$: The disparity histogram in the pursuit mode based on the former target mask provides the current target disparity, $d_p = 13$, while that in the saccade mode based on the newly detected moving target provides the disparity of the new target candidate[2], $d_s = 13$. Since $d_s = d_p$, the new candidate is no closer than the current target, and the pursuit target mask is selected as the final target mask.

Frame k: Pursuit and saccade disparities in this frame are $d_p = 13$ (target person staying at the same depth) and $d_s = 12$ (the second person approaching). Since $d_s < d_p$, the saccade target mask is selected, that is, attentional shift takes place. Notice that the former target mask $T(k-1)$ is fed back.

Frame $k+1$: Analogously $d_p = 12$ and $d_s = 13$ as the attention has been shifted in the previous frame. The pursuit target mask is selected since $d_s > d_p$ and the attention stays on the second person.

The above processes are conducted in each frame to determine the target mask, which continuously provide the clue to the attentional target throughout the sequence of images.

5 Conclusion

We have presented a computational model for visual attention with an early parallel stage with preattentive cues followed by a later serial stage where the cues are integrated. Key points in our model are:

- the use of multiple cues, i.e. stereo, image flow and motion,

[2]The small peak observed beside the main peak in saccade mode is arising from the leg of the table that is incidentally detected by the motion detection module.

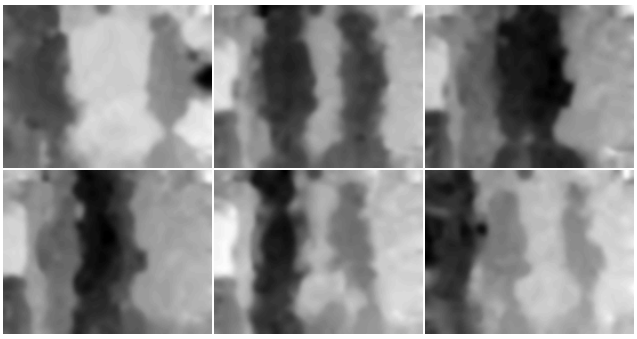


Figure 6. Disparity maps computed for the sequence in Figure 5. The darker, the closer.

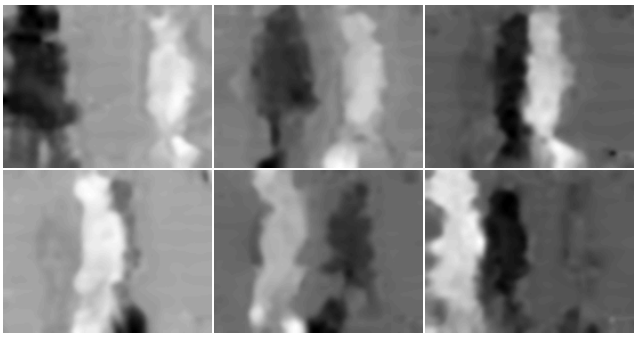


Figure 7. Horizontal fbw maps computed for the sequence in Figure 5. The lighter, the more leftward.

Figure 8. Detected motion for the sequence in Figure 5. The darker, the stronger.

Figure 9. Target masks computed for the sequence in Figure 5.

Figure 10. An example sequence with 2 moving persons taken by a rotating binocular camera head. Every 10th frame of the left image is shown (40 msec between frames).

Figure 11. Detected motion for the sequence in Figure 10. The darker, the stronger.

Figure 12. Target masks computed for the sequence in Figure 10.

- the use of relative depth as a target selection criterion,
- that the model can be implemented by simple computations.

We have shown experimentally that the model provides expected results for a given control scheme for target selection based on nearness and motion. In particular this also demonstrates that sufficient information for our model is computable by simple algorithms. Our model therefore shows promise as a basis for investigating the "where to look next" problem more generally.

Acknowledgment

This work benefited greatly from discussions with the vision group CVAP at KTH. Comments from Tomas Uhlin and Jonas Gårding are particularly acknowledged. This work has been supported by TFR, the Swedish Research Council for Engineering Sciences, which is gratefully acknowledged.

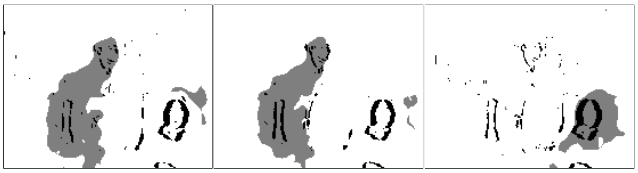
References

[1] A. Maki. *Stereo Vision in Attentive Scene Analysis*. Ph. D. dissertation, Royal Inst. Tech., 1996. ISRN KTH/NA/P--96/07--SE.

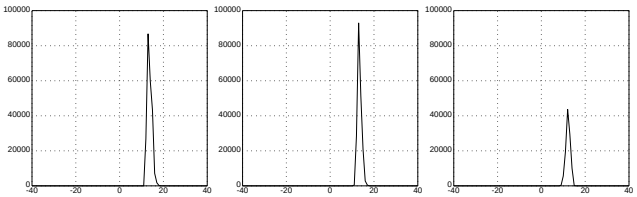
[2] A. Maki and T. Uhlin. Disparity selection in binocular pursuit. *IEICE, the Institute of Electronics Information and Communication Engineers Transactions on Information and Systems*, E78-D(12):1591–1597, December 1995.

[3] D. W. Murray, K. J. Bradshaw, P. F. Mclauchlan, I. D. Reid, and P. M. Sharkey. Driving saccade to pursuit using image motion. *IJCV*, 16:205–228, 1995.

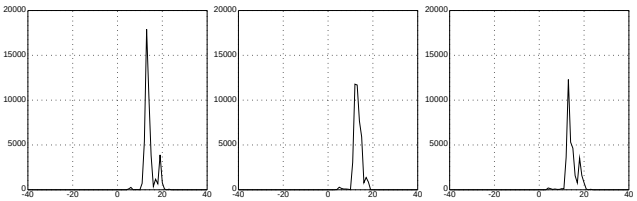
[4] K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320(20):264–265, March 1986.



(a) Detected motion masks (dark) and target masks in the former frame (gray).



(b) Disparity histograms in pursuit mode based on the former target mask.



(c) Disparity histograms in saccade mode based on newly detected motion.

(d) The resulting target masks superimposed on the original image sequence.

Figure 13. The process producing the target masks in frame $k-1, k, k+1$ (left, middle, right). The horizontal and vertical axes in the histograms are for disparity estimates and sum of corresponding certainty values.

[5] P. Nordlund and T. Uhlin. Closing the loop: Pursuing a moving object by a moving observer. Technical Report ISRN KTH/NA/P--95/06--SE, Royal Inst. Tech., 1995. Shortened version in Proc. 6th International Conf. on Computer Analysis of Images and Patterns. Also in IVC vol. 14, no 4, May 1996, pp 265–275.

[6] K. Pahlavan and J.-O. Eklundh. A head-eye system — analysis and design. *CVGIP: IU*, 56(1):41–56, 1992.

[7] T. F. Syeda-Mahmood. Data and model-driven selection using color. In *2nd ECCV*, pages 115–123, May 1992.

[8] A. Treisman. Preattentive processing in vision. *CVGIP*, 31:156–177, 1985.

[9] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, N. D. Y. Lai, and F. Nufb. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

[10] T. Uhlin, P. Nordlund, A. Maki, and J.-O. Eklundh. Towards an active visual observer. In *5th ICCV*, pages 679–686, June 1995.

[11] J. M. Wolfe and K. R. Cave. Deploying visual attention: The guided search model. In A. Blake and T. Troscianko, editors, *AI and the Eye*, chapter 4, pages 79–103. John Wiley & Sons Ltd., 1990.