Attentional Scene Segmentation: Integrating Depth and Motion from Phase

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

Abstract *We present an approach to attention in active computer vision. 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 [26, 30] 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 [25], with an early parallel stage with preattentive cues followed by a later serial stage where the cues are integrated. The contributions in our approach are (i) the incorporation of depth information from stereopsis, (ii) the simple implementation of low level modules such as disparity and flow by local phase, and (iii) the cue integration along pursuit and saccade mode that allows us a proper target selection based on nearness and motion. We demonstrate the technique by experiments in which a moving observer selectively masks out different moving objects in real scenes.*

Keywords: *Attention, stereoscopic depth, image flow, motion, cue integration, pursuit, saccade, target selection*

*This report describes research done while the author was with CVAP, the Royal Institute of Technology (KTH). Present address: TOSHIBA Kansai Research Laboratories 8-6-26 Motoyama-Minami-Cho, Higashinada-ku, Kobe, 658 Japan.

2 Introduction

Any complex visual system will need a good selection mechanism to allow it to allocate its spatially limited processing resources appropriately [30]. Attention is a mechanism that enables vision systems to cope with the tradeoff between the amount of information to be processed and the substance of the process itself. Designs of computer vision systems or algorithms are often motivated by the biological counterparts one way or another. This can be attributed to the fact that biological systems have in their evolution acquired functions which actually work in the real world. An example of such a function which is elegantly achieved in biological systems is the attentional mechanism. It serves to efficiently reduce the enormous amounts of available information such that it can be selectively processed. The main theme of this paper is to present a computational approach to such attentional mechanisms. The emerging question is how to achieve such mechanisms and what kind of criterion to employ out of the enormous amount of basic features that may be observed in a scene. These basic features include for example color, orientation, size, motion, and stereoscopic disparity. In designing a computational framework of such attention mechanisms and in choosing among these basic features, we believe that it is worth while to refer to the successfully functioning human attentional system. Nakayama and Silverman [17] state in their reports on psychophysical observations:

> "We speculate that retinal disparity in addition to retinal locus has priority when compared with other visual stimulus dimensions..." (p. 265)

Retinal disparity is defined as the retinal displacement between the two projections of an object on the left and right retina. In the human visual system binocular stereopsis, by way of retinal disparity, provides an important means for depth perception. Furthermore nearness in depth is directly connected to urgency in spatio-temporal environments. With the motivation that this should also provide one of the strongest cues in computer and machine vision, we employ binocular stereopsis as the central cue for our computational approach to attention.

In this work, as well as stereopsis, we propose to base the system also on image flow and motion in its framework. As seen in the schematic diagram of our framework in Figure 1, in this scheme cue integration and attention over time are essential aspects. Part of the cue integration work has appeared in [27]. The contribution here is that we show that the system can attend to different targets in a purposive way in a cluttered environment. The second key point in this context is the use of depth information, as suggested is done in human vision by Nakayama and Silverman [17]. The computation of precise depth information is generally a time consuming task. The 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, i.e., we propose to employ local phase information to derive both depth and flow. 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. The important point to note is that the required information is indeed computable and that the desirable behavior of the system is acquired.

The organization of the paper is as follows. We first make a brief overview of relevant issues on the human attentional system based on psychophysical reports in Section 3. Section 4 then introduces some of the earlier works on attention. Describing the low level modules in Section 5, we design the cue integrations along pursuit and saccade mode in Section 6. Section 7 exemplifies the performance of the proposed prototype through experiments, and Section 8 finally concludes the paper.

3 Observations about human attention

The notion of attention plays an important role in biological vision in terms of selecting a part of the scene out of the massive flow of information in space and time. Posner and Petersen [21]

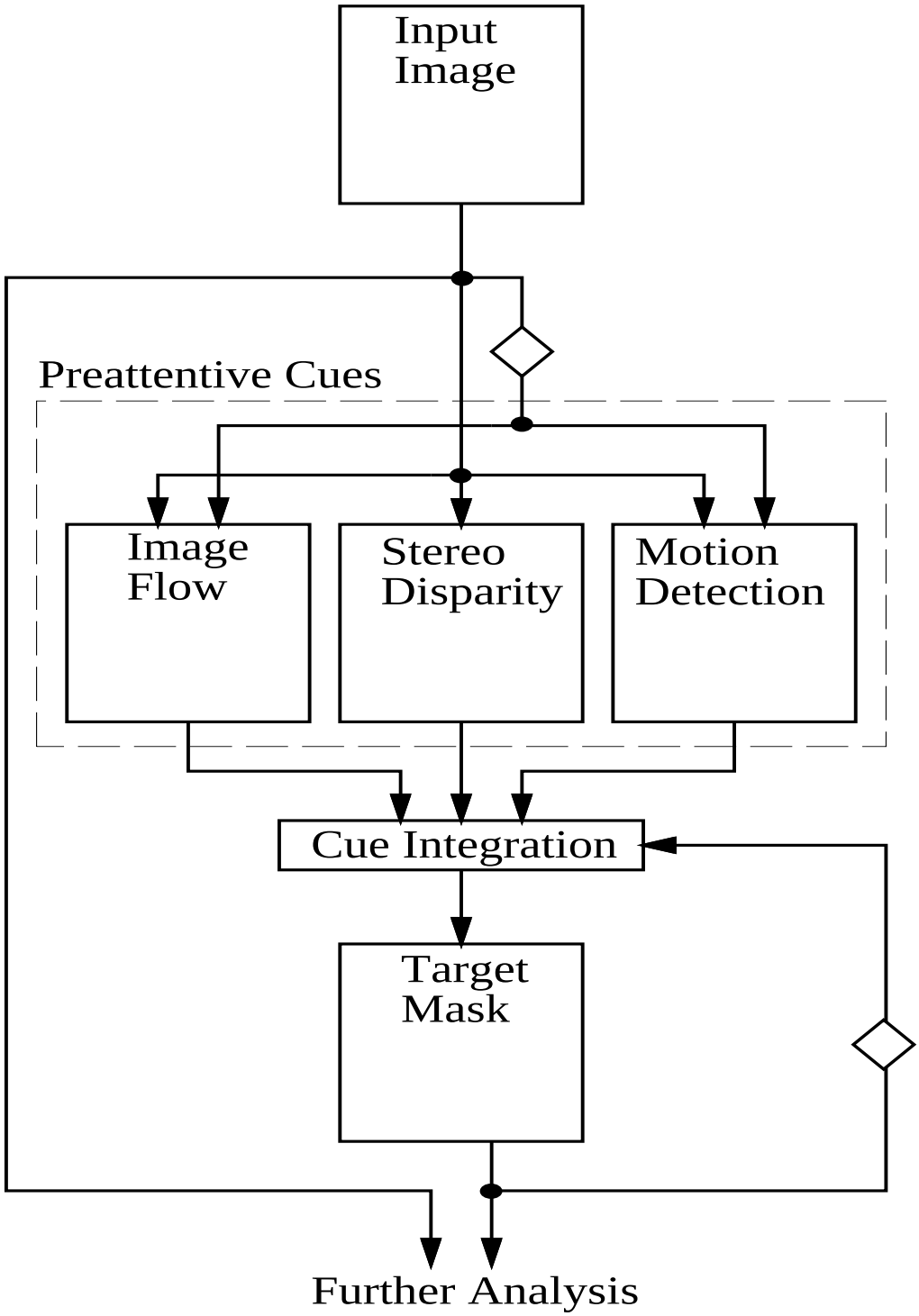


Figure 1: *A schematic diagram of the proposed attentional framework. It follows the general concept of visual attention, i.e., the early parallel stage with preattentive cues and the later serial stage where the cues are integrated. The diamonds indicate a one frame delay.*

discuss the attentional system of the human brain by dividing it into three major subsystems that perform different but interrelated functions, these are (i) orienting to sensory events, (ii) detecting signals for focal processing, and (iii) maintaining an alert state. Among those functions, we will in this paper mostly be concerned with orienting to sensory events, including issues on visual locations and search.

Visual locations: Visual orienting is usually defined in terms of the foveation of a stimulus (overt by physically directing attention). Foveating a stimulus improves efficiency of processing targets in terms of acuity, but it is also possible to change the priority given a stimulus by attending to its location covertly without any change in eye position [21]. In other words, it is almost as if we have some "internal spotlight" that can be aligned with an object to enable us to see it [8]. Coren et al. [6] use the metaphor "attentional gaze" to conceptualize some of the findings in covert visual orienting and visual search, and describe it like:

> "Covert shifts in the attentional gaze seem to behave in a similar way to physical movements of the eye and attentional gaze usually cannot be drawn to more than one location in the visual field at any instance in time. Experimental studies of the attentional gaze show that it can shift much faster than the eye — it reaches a stimulus location before the eye does and seems to help to guide the eye to the proper location." (Coren et al. 1993, p. 517)

They also indicate three aspects of the attentional gaze which are important in the processing of sensory information at any one moment, i.e. a locus, an extent, and a detail set [6]. Attentional gaze poses an important issue on active vision [1, 2, 3], in the sense that it provides a clue to guide the gaze direction spatio-temporally. Our attentional scheme appearing later treats to the aspects of locus and extent while it does not specify any particular detailed set.

Visual search: Starting from a neurophysiological viewpoint, all neurons are selective in the range of activation to which they will respond. The role of the attention system is to modulate this selection for those types of stimuli that might be most important at a given moment. To understand how this form of modulation operates, it is important to know how a stimulus would be processed without the special effects of attention [21]. In cognition, unattended processing is called "automatic" to distinguish it from the special processing that becomes available with attention. It is also termed "preattentive processing" in dichotomy with attention.

In her article on preattentive processing from a psychological viewpoint, Treisman [25] poses two different problems as follows:

> "One is to define which features or properties are the basic elements or visual primitives in early vision. The second concerns how they are put together again into the correct combinations to form the coherent world that we perceive." (Treisman 1985, p. 157)

She performs experiments focusing on texture segregation and visual search, where subjects are asked to look for a particular target in displays containing varying numbers of distractor items. If the target is defined by a simple visual feature, detection appears to result from parallel processing; the target "pops out" and the search time is independent of how many distractors surround it. Such experiments turn out to be consistent with the idea that early visual analysis results in separate maps for separate properties, and that these maps pool their activity across locations, allowing rapid access to information about the presence of a target, when it is the only item characterized by a particular preattentively detectable feature [25]. Though there is not complete agreement about the set of basic features, there is agreement about many members of the set, such as color, orientation, size, motion and stereoscopic depth [30].

Also reported is that the search for an item defined by the conjunction of two stimulus dimensions is conducted serially, and the search time increases as the number of items becomes larger.

Thus, it seems that the visual system is incapable of conducting a parallel search over two stimulus dimensions simultaneously. Nakayama and Silverman [17] extend this conclusion for the conjunction of motion and color. Interestingly, they also point out two exceptions through similar experiments: if one of the dimensions in a conjunctive search is stereoscopic disparity, a second dimension of either color or motion can be searched in parallel.

Wolfe and Cave [30] review evidence of preattentive processing in the human visual system through experiments and put it in a nice way:

> "...the human visual system uses fast, efficient, parallel algorithms to guide the later serial stages of visual processing. This attentional system can use information about expected stimuli when it is available, but will identify unusual parts of the input whether or not they are expected. These parallel mechanisms are very limited, both in their accuracy and in what they can do, but by guiding the allocation of higher-level, limited-capacity processes, they can speed visual processing considerably." (Wolfe and Cave 1990, p. 102)

The intension here is to employ similar simple and efficient mechanisms to obtain attentional algorithms in a computer vision system to quickly identify those parts of the input that should be processed immediately. Particularly we will try to obtain the observed advantage of stereoscopic disparity as a basic feature which can be processed in parallel with some other basic feature as seen in [17].

4 Related work

A number of computational approaches of attention have been proposed on the basis of the psychophysical findings on human visual attention. Attentional mechanism can be highly task-dependent and different approaches take different modalities while sharing certain concepts such as the division into preattentive and attentive processing stages.

Burt [4] describes attention from a computer vision perspective by three elements: foveation, tracking and high level interpretation. A rudimentary fovea is formed within a Laplacian pyramid and tracking is performed to isolate selected regions of a scene in time, by canceling computed background motion. He defines an object representation called a *pattern tree* for a fast, hierarchical structured search arguing that very fast reasoning processes are needed to interpret partial scene information as it is gathered, and to provide informed guidance for where to look next. The pattern tree description of objects known to the system are sorted in a knowledge base and the information is gathered in a task oriented fashion.

An attentional model applying the notion of winner-take-all is presented by Clark and Ferrier [5]. The first stage in the model extracts *primitives* in parallel across the visual field. The results from this stage are a set of *feature maps* which indicate the presence or absence of a feature at each location in the image. The next stage of the model combines the results from the feature maps. The output from the feature maps are amplified with different gains for each map and then summed to form the *saliency map*. Finding the location with the maximum value gives the most salient location with respect to the given amplifier gains, which may vary over time, thus changing the location of the most salient feature. However, only one location can be attended to at one time. The employed features are blobs, moments of objects, and the intensity value.

Sandon [22] bases his attentional vision system on feature and saliency maps as well, but aims for recognition of an object in a particular location in an image. The attentional system consists of a feature processor, an object processor and an attention processor. According to the needs of the current task, feature maps extracted from the image by the feature processor are gated by the object processor, which maintains a set of object models, to provide inputs to the attention processor. The attention processor combines the activity from the feature maps to compute the saliency map, which represents the importance of each region of the image. According to the result of attentional competition, in addition, it gates regions of feature maps to provide inputs to

the object processor. A feature called an *edge histogram*, including edge information (magnitude and orientation), is used for the feature maps.

The work by Syeda-Mahmood [24] is also based on object models. The problem domain is defined within the scope of model-based object recognition, and the experiments are limited to static scenes. The idea is to determine which regions in an image are likely to come from a single object, and color is employed as the feature for this segmentation.

In the approach proposed by Westelius [28], edge information and rotation symmetry are adopted as the features to form a potential field for driving attention. Phase information from quadrature filters is used to generate a potential field drawing the attention towards and along lines and edges in the image. To find objects another potential field is generated using rotation symmetries. The potential fields are weighted together differently depending on the state of the system, and the state is in turn determined by the type and quality of the data in the fixation point. The states are, (i) search line, (ii) track line, (iii) avoid object, and (iv) locate object.

An attentional prototype for early vision is developed by Tsotsos et al. [26] in which they emphasize the neurobiological plausibility of it. The model consists of a set of hierarchical computations, involving the idea of an *inhibitory beam* which expands as it traverses through the entire hierarchy but is kept to a limited size by inhibition. Any salience map can be used as the stimulus at the input level, and the areas of attention are acquired as selected *receptive fields* at the bottom level.

Our work has in several ways been inspired by the described efforts. An essential additional contribution of ours is the incorporation of three dimensional information from stereopsis [12]. In particular, we use stereoscopic disparity derived from the fact that two eyes receive slightly different views of the three-dimensional world.

Concerning the schemes for measuring disparity as a static depth cue in stereo vision, a wealth of algorithms has been suggested. Most of the techniques fall in one of the two categories of area-based correlation and feature-based correspondence, and those approaches to stereopsis are in a sense complementary [9]. Correlation produces a dense set of responses but has difficulty with constant or rapidly changing structure and with interocular image difference, while feature-based correspondence avoids some of these problems by considering the image at different scales but then fails to obtain a dense set of responses by matching only sparse tokens. It would be desirable both to avoid the problem of structure at different scales and at the same time to produce a dense response. Some additional techniques have been developed to complement some of the shortcomings, such as multiple-baseline stereo [19], non-parametric local census transform [31], or use of linear spatial filters tuned to a range of orientations and scales [11]. It is on the other hand common in both techniques that a search process is required to find the best match between the areas or features.

As a third approach, a new technique has been proposed in which disparity is expressed in terms of phase differences in the output of local, bandpass filters applied to the stereo views [10, 23, 29]. The main advantages of such local, phase-based approaches include computational simplicity, stability against varying lighting condition and especially direct localization of the estimated disparities. Also, there is biological evidence supporting different aspects of this method including [23], e.g. bandpass spatial-frequency filters are thought to be plausible models of the way in which the primary visual cortex processes a visual image [7].

We consider the characteristics of this search-free and thus fast approach as suitable to provide stereoscopic disparity as a basic feature in preattentional early vision. We therefore employ the technique in our computational approach to attention while the traditional alternatives appear to be less well suited in such an application.

5 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.

5.1 Stereo disparity

Relative depth, that plays a central role in our system, 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 is briefly introduced in Appendix A.1. The employed multi-scaled scheme based on the algorithm is described in [15]. 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 Appendix A.2 and [14] for details of the procedure producing the target mask. A point to be noted is that the resulting 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.

5.2 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 [16]. 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.

5.3 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. The technique is formulated in Appendix **??** and full description is found in [18].

6 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 as sketched in Figure 1. 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 [13] and duration-based one are considered.

6.1 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 sequence 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.

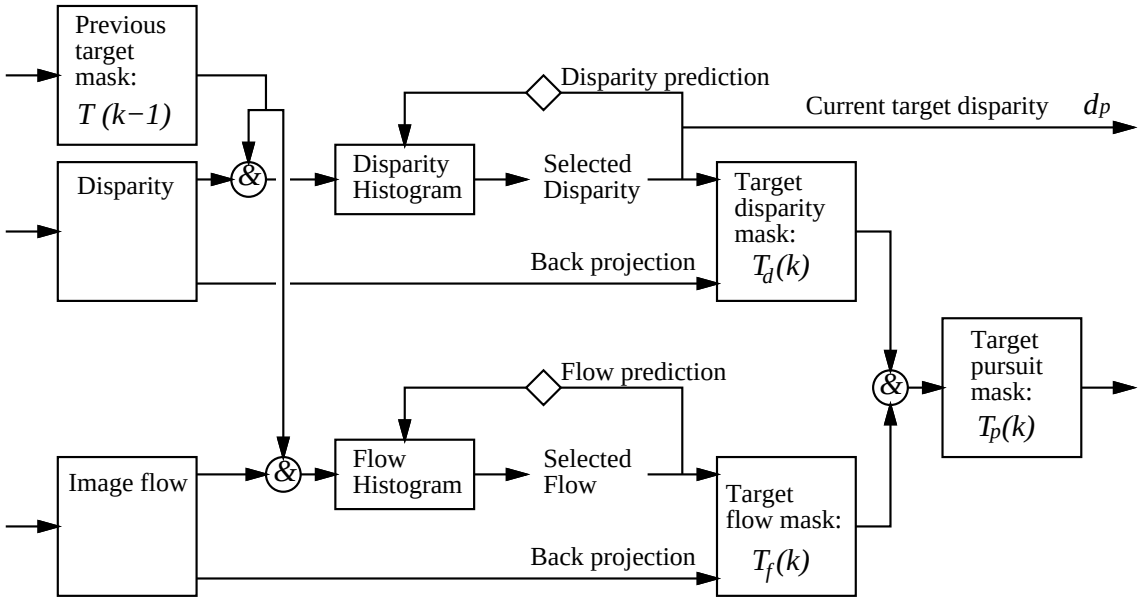


Figure 2: *Schematic flow diagram of the attentional pursuit. It composes part of the "cue integration" in the framework shown in Figure 1. The diamonds indicate a one frame delay in the feedback. The circles with & indicate a logical and operation.*

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 that is consistent with both disparity and flow remains. The process in frame k can be formulated as:

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

6.2 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

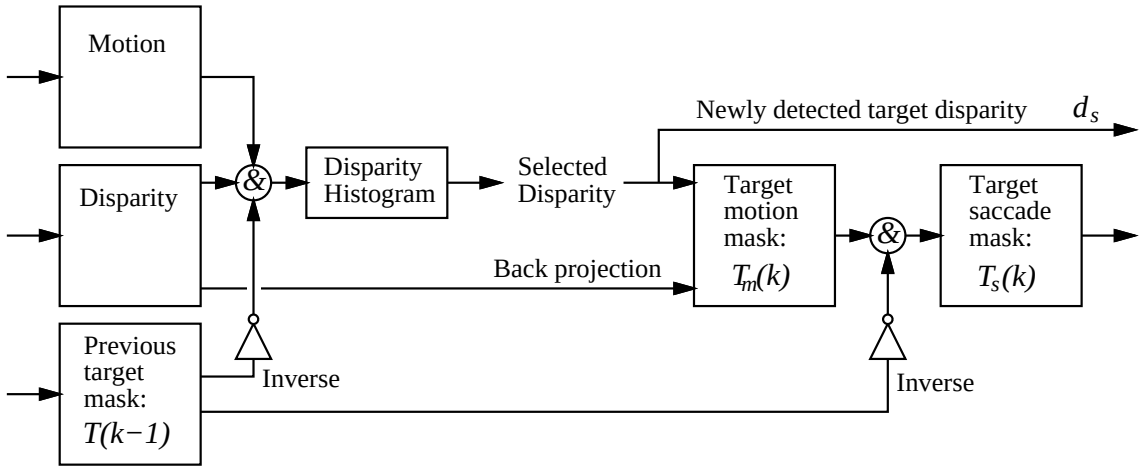


Figure 3: *Schematic flow diagram of the attentional saccade. It composes part of the "cue integration" in the framework shown in Figure 1. The circles with & indicate a logical and operation.*

target. The process is completed by inhibiting the produced target mask again by $T(k-1)$ to confirm 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.

6.3 Target selection

The cue integration process described thus 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). Two different criteria are proposed in the following, i.e., depth-based criterion and time transient criterion.

Depth-based criterion While the framework introduced is open to accept different criteria, we have mainly 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 who 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. Thus, the closest moving object is kept on attended to over time.

$$\begin{aligned} T(k) \quad &= \quad \begin{cases} T_s(k), & \text{for } d_s \leq d_p \\ T_p(k), & \text{otherwise} \end{cases} \\ d_s \quad &: \quad \text{Disparity of newly detected target} \\ d_p \quad &: \quad \text{Disparity of the current target} \end{aligned} \tag{3}$$

Duration-based criterion Certain tasks, such as recognition, identification or computing motion may require that the system attends to the target for some certain time instance so that the relevant process can produce the desired result. In addition to the depth-based criterion, thus, we have considered a duration-based criterion. With this criterion, attention to some object is kept on

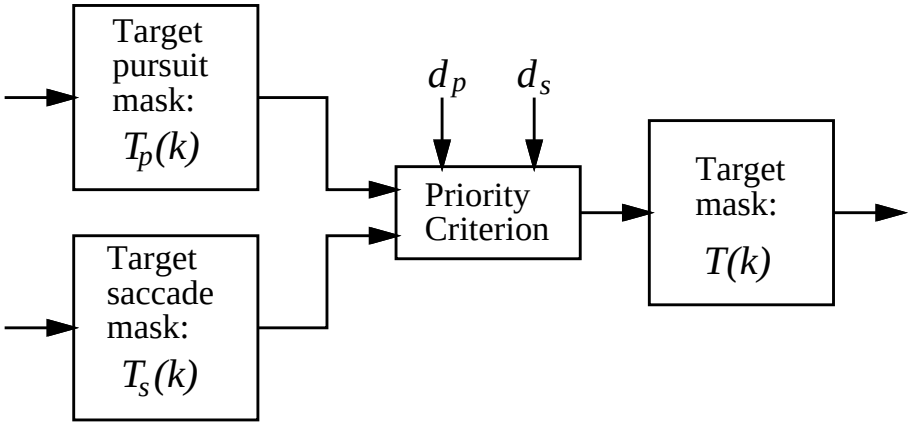


Figure 4: *Attentional target. In each frame either target pursuit or target saccade mask is selected as the final target mask according to a criterion. For example in the depth-based criterion the mask covering the closer target is selected on the basis of the comparison between the disparity of the current target d_p and that of newly detected target d_s.*

for a predefined duration and then shifted to some other object, according to the depth-based criterion as soon as any candidate appears. Ideally the time instance the system attends to the target object should be task dependent. Although our scheme has so far not been integrated with other processes, experiments are performed assuming some certain time duration.

7 Experiments

The above described attentional scheme has been examined through experiments. In the experiments we purposively choose to include humans as objects in the scene. The humans may be either stationary or moving and may even change their course of travel unpredictably. As such they provide reasonable attention candidates in a dynamic world. Moreover, in view of the circumstances in which a vision system should function, humans are common and important targets to relate and react to, and thus to attend to. Humans also form realistic prototype objects in for example obstacle avoidance, as obstacles to avoid may be either moving or stationary.

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

The resulting target masks according to the depth-based criterion are shown in Figure 10. The frames are numbered from 1 to 8. It is observed that the target masks are produced in consistency with each of the disparity maps, the certainty maps and the horizontal flow maps. Figure 11 demonstrates the histograms for disparity selection in the corresponding frames. Histograms in the pursuit mode based on the former target masks are indicated in lines, whereas those in the saccade mode based on newly detected motion are in dashed lines. Appearing peaks represent attended targets or target candidates, and it is observed that the attention is taken over from one object to the other depending on the relative depth information. A detailed description of the performance of the system along the frames is given in the following.

[1]The gray scale in the displayed 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 1-3: By nearness and motion the system first picks out the person A, walking from the left towards the center. Two peaks are appearing in the histograms at the corresponding frames. The person A is represented by the peaks in lines as the current target. The peak shifts as the person A moves further away. Note that smaller disparities correspond to closer objects in depth. Peaks in dashed lines on the other hand represent another moving person B, passing from the right hand side. In Frame 3 the person A and B are almost in the same distance and the two peaks are overlapping. The third person C in behind is not appearing in the histogram because of no motion.

Frame 4-6: At Frame 4 the attention to the person A is replaced with the second person B who is then relatively closer than the person A. The person B is in attention while being present until Frame 6, and represented by the peaks in line. In the histograms, in the meantime, two peaks in dashed line are observed for the saccade mode by the person B and person C. They are once merged into one peak at Frame 11 when two persons are in the same depth (see Figure 6).

Frame 7-8: The attention is finally shifted to the third person C when B disappears from the scene at Frame 13. Histograms in the saccade mode are vanishing as no object except for the attended one is in noticeable motion.

Throughout the sequence the system basically masks one target at each frame following the depth-based criterion. In situations when several objects belong to an overlapping range both in depth and flow, however, more than one target could accidentally be masked. For example in the Frame 7 in the sequence in Figure 10, it is seen that parts of the person A and the table on the left corner are masked as well as the target person C. While this is largely a matter of defining the range of target depth and flow, the result is natural because objects with completely identical disparity and flow would not be recognized separately. The situation also indicates the possibility to improve the scheme by merging the system with some extra cues such as information about location in the early parallel process.

Figure 12 shows another sample image sequence, this time by moving cameras [20]. 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, passing by in front and disappearing on the left end, while the observing camera head is moving laterally. Every 10th frame is shown (images are taken at framerate 25 Hz). Figure 13 shows that the motion detection process functions even for a sequence with moving cameras. The resulting target masks are shown in Figure 14, where it is observed that the closest moving object is kept attended to.

Figure 15 exemplifies the process of the moment when the target mask is shifted from one target to the next one. Illustrated are the masks restricting the input to the disparity histogram, two histograms in pursuit and saccade modes, and 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.

Frame $k-1$: The disparity histogram in pursuit mode based on the former target mask provides the current target disparity, $d_p = 13$, while that in saccade mode based on the newly detected moving target provides the disparity of the new target candidate, $d_s = 13$. 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 closer). Since $d_s < d_p$, the saccade target mask is selected, i.e., attentional shift takes place. Notice that the former target mask $T(k-1)$ is fed back.

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

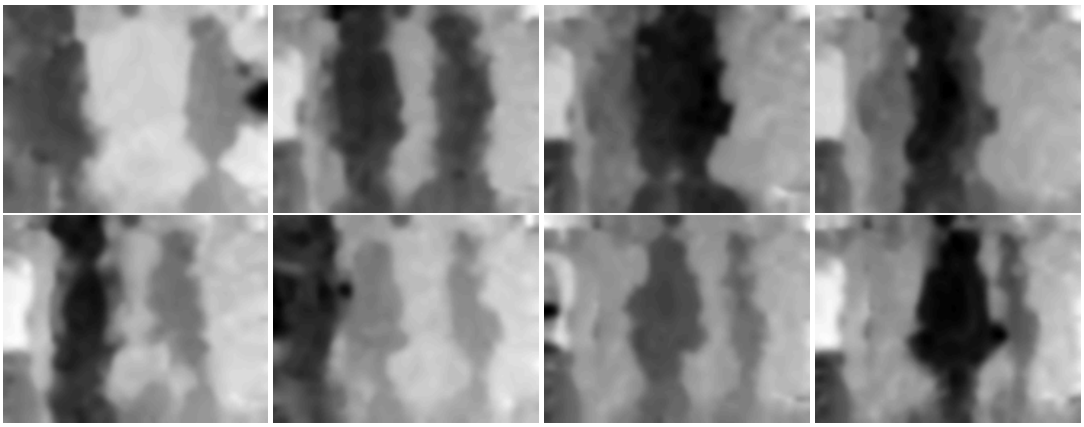


Figure 6: *Disparity maps computed for the 3 persons sequence. The darker, the closer.*

Figure 7: *Certainty maps computed corresponding to the disparity map of the 3 three persons sequence. The lighter, the higher certainty.*

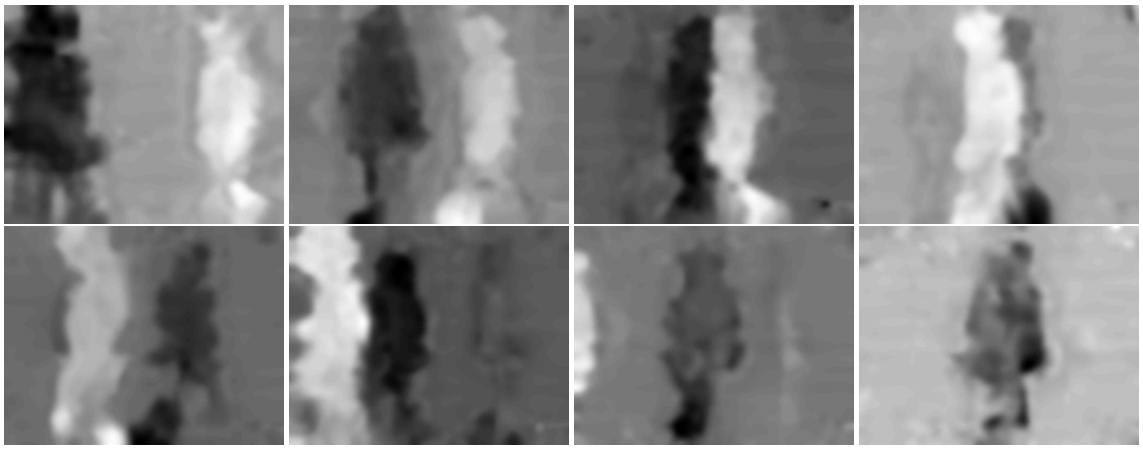


Figure 8: *Horizontal flow maps computed for the 3 three persons sequence. The lighter, the more leftward.*

Figure 9: *Detected motion for the 3 three persons sequence. The darker, the stronger.*

Figure 10: *Target masks computed for the 3 three persons sequence.*

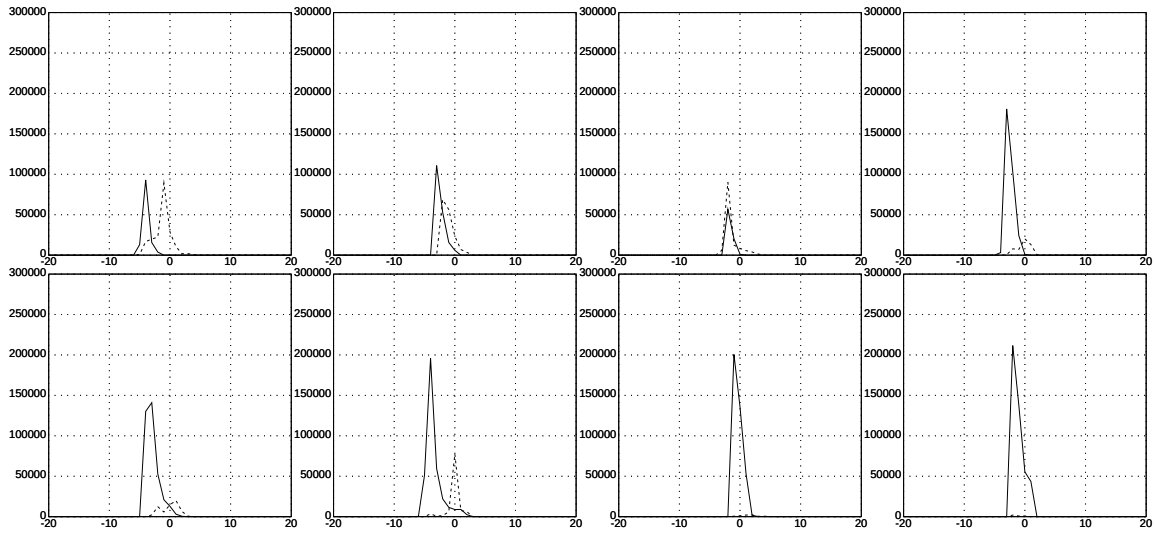


Figure 11: *The disparity histograms in the pursuit mode based on the former target mask (solid line) and in the saccade mode based on newly detected motion (dashed line). The horizontal and vertical axes are for the disparity estimates and the sum of corresponding certainty measures. Small disparities mean closeness in depth.*

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

Continuous processes such as above are conducted to determine the target mask in each frame, providing the clue to the attentional target throughout sequence of images.

As the final experiment the duration-based criterion is examined using a sequence shown in Figure 16. It includes again the 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 17 the target masks are shown. The frames are numbered from 1 to 12. It is demonstrated that the compulsory attentional shift occurs periodically by the duration-based criterion (duration is fixed to 10 frames). A detailed description of the attentional shift is given for each period of the sequence.

Frame 1-3: 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 takes on sufficient movement to be attended to.

Frame 4-5: The attention is shifted to C whose movement begins to be recognized. For the fixed duration (10 frames), C is kept as the target in the pursuit mode.

Frame 6-11: The attention is shifted back and forth while keeping the fixed duration. First back to A, then to C and again to A. During this period B with little movement is not among the candidates for the attention.

Frame 12: This time the attention is shifted instead to B that is finally starting to move in closer distance than C is. The depth-based criterion is reflected here while employing the framework of the duration-based criterion.

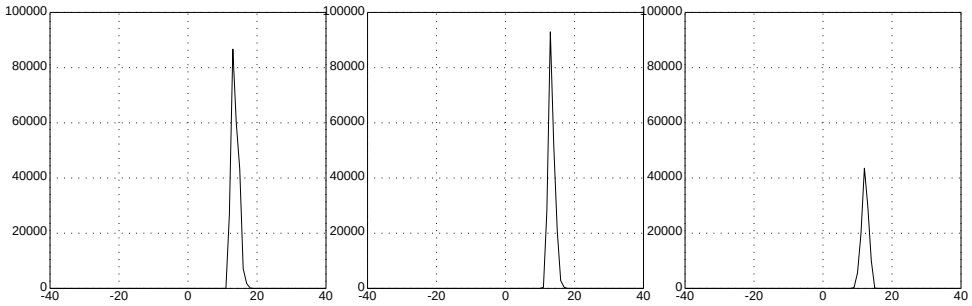
Overall it is seen that the behavior of the system is following the duration-based criterion. Though the duration is fixed to 10 frames here, the choice is arbitrary. It should be determined depending on the consecutive process on the attended target. There is some noise observed in masks in

Figure 12: *An example sequence with 2 moving persons taken by a moving binocular camera head. Top-left to bottom-right. Every 10th frame of the left image is shown (40 msec between frames).*

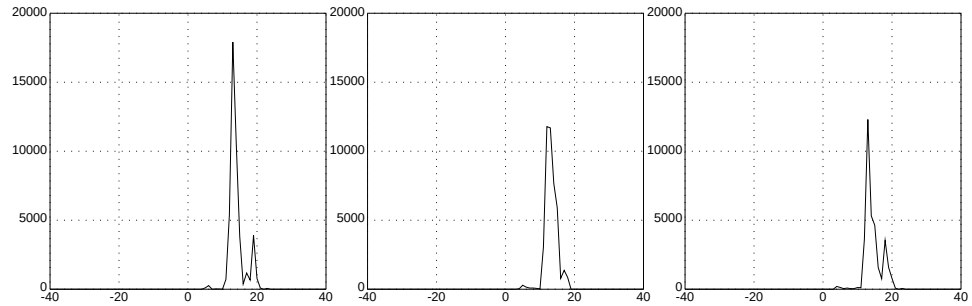
Figure 13: *Detected motion for the 2 persons sequence. The darker, the stronger.*

Figure 14: *Target masks computed for the 2 persons sequence.*

(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 15: *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.*

certain frames such as Frame 4 and 8. It is mainly related to the fact that ranges of the disparity and horizontal flow is fixed here for the segmentation of a target. That is, by a fixed range of disparity different depth intervals are covered depending on the distance: the further, the larger interval. A fixed range of flow also corresponds to varying displacement in time due to the motion parallax. The employment of a flexible range definition considering the features of the attended targets would improve the performance, and is an issue for further investigation.

8 Conclusion

In this paper we have proposed a computational approach to visual attention. The specific character of the work presented here is in the purposive target selection. Considering the problem of attention using image sequences over time, we have specified the attended part of input image by providing a target mask in each frame. Segmented disparity map based on histogramming provides the central cue in the proposed system, in cooperation with information of image flow and a motion detection technique. With the early parallel stage of preattentive cues, the consecutive stage integrated those cues. Key points in our system are summarized as:

- the employment of relative depth as a target selection criterion as suggested is performed in human vision,
- the simple computation of low level cues such as disparity and flow by local phase,
- the integration of cues along pursuit and saccade mode realizing the purposive target selection.

We have shown experimentally that the system 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 system is computable by simple algorithms. The proposed approach to visual attention 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.

Figure 16: *An example sequence with 3 persons taken by a stationary binocular camera head. Top-left to bottom-right. Every 5th frame of the left image is shown (40 msec between frames).*

Figure 17: *Target masks computed for the sequence in Figure 16. Top-left to bottom-right. Every 5th frame is shown (40 msec between frames).*

A Appendix

A.1 Phase-based algorithm

Disparity map The fundamental idea of the phase-based approach to disparity estimation is to recover local disparity as the spatial shift from the local phase difference observed in the Fourier domain. In practice, the phase is extracted by taking the argument of the convolution product $V_l(\mathbf{x})$ and $V_r(\mathbf{x})$, which is produced at each point $\mathbf{x}$ in the image by convolving a complex filter[2] with the left and the right stereo images[3]. As the local shift between stereo images is approximately proportional to the local phase difference, a disparity estimate at each point in the image is derived accordingly:

$$D(\mathbf{x}) \quad \simeq \quad [\arg V_l(\mathbf{x}) - \arg V_r(\mathbf{x})]/\omega(\mathbf{x}). \tag{4}$$

$D(\mathbf{x})$ denotes disparity at $\mathbf{x}$ and $\omega(\mathbf{x})$ represents some measure of underlying frequency of the image intensity function in the neighborhood of $\mathbf{x}$, which in this case is computed as phase derivative. For details on the techniques employed here, see [12, 15].

Certainty map In order to check the feasibility of the estimated disparity and threshold unreliable estimation, we also compute a certainty value C($\mathbf{x}$) defined on the basis of the magnitude of the convolution product:

$$C(\mathbf{x}) \equiv \sqrt{|V_l|\,|V_r|} \cdot \frac{2\sqrt{|V_l|\,|V_r|}}{|V_l| + |V_r|}. \tag{5}$$

A.2 Disparity selection

Disparity-certainty histogram Based on a disparity map $D(\mathbf{x})$ and a certainty map $C(\mathbf{x})$, we compute a histogram $H(D_d)$ with respect to the discrete disparities D_d:

$$H(D_d) = \sum_{\mathbf{x}} C(\mathbf{x}) \;\text{ for }\; \{(\mathbf{x}) \mid D_d \leq D(\mathbf{x}) < D_{d+1}\}. \tag{6}$$

$H(D_d)$ is defined as the sum of the certainty values at the pixels where the disparity is estimated to D_d. Multiple peaks appear in the histogram corresponding to objects with different disparities.

Disparity prediction With a prediction of what disparity the target should have, the closest peak in the histogram can be selected as the estimate of the target disparity. For computational simplicity a linear predictor is used with a weighting factor α[4]:

$$D_p(k+1) \quad = \quad D_s(k) + P(k) \tag{7}$$

$$P(k) \quad = \quad \alpha \cdot (D_s(k) - D_s(k-1)) + (1-\alpha) \cdot P(k-1) \tag{8}$$

where $D_s(k)$ and $D_p(k)$ represent the selected and predicted disparity at frame number k while $P(k)$ denotes the predicted change.

A.3 Brightness constancy and affine image velocity

Brightness constancy Using the notation $I_x = \frac{\partial I(\mathbf{x},t)}{\partial x}, I_y = \frac{\partial I(\mathbf{x},t)}{\partial y}, I_t = \frac{\partial I(\mathbf{x},t)}{\partial t}$ and $\nabla I(\mathbf{x},t) = (I_x(\mathbf{x},t), I_y(\mathbf{x},t))^T$, the brightness constancy can be written as,

$$\nabla I(\mathbf{x},t) \cdot \mathbf{v}(\mathbf{x},t) + I_t(\mathbf{x},t) = 0 \tag{9}$$

where $\mathbf{v} = (\frac{dx}{dt}, \frac{dy}{dt})^T$ is the image velocity. This equation is not enough to constrain the two parameters of $\mathbf{v}$, given the gradients (∇I), and the time derivatives (I_t). What can be determined though, is $\mathbf{v}$'s component normal to the gradient, the normal image velocity.

[2] We employ discrete approximations to the first and second derivatives because of their computational simplicity.
[3] Two consecutive images are used instead when it is to derive horizontal image flow.
[4] $0 \leq \alpha \leq 1$, in the experiments α is set to be 0.2 to attenuate the influence of noise.

Residual normal velocity calculation From (9) we can determine the normal component of the velocity vector locally as:

$$v_n(\mathbf{x}) = -\frac{I_t(\mathbf{x})}{|\nabla I(\mathbf{x})|}. \tag{10}$$

With an arbitrary velocity field, $\hat{\mathbf{v}}(\mathbf{x})$, and the gradients in an image, we can write the normal velocity as:

$$\hat{v}_n(\mathbf{x}) = \frac{\nabla I(\mathbf{x}) \cdot \hat{\mathbf{v}}(\mathbf{x})}{|\nabla I(\mathbf{x})|}, \tag{11}$$

and define the residual between this and the observed normal velocity as:

$$R(\mathbf{x}) = \hat{v}(\mathbf{x}) - v_n(\mathbf{x}) = \frac{\nabla I(\mathbf{x}) \cdot \hat{\mathbf{v}}(\mathbf{x}) + I_t(\mathbf{x})}{|\nabla I(\mathbf{x})|} \tag{12}$$

The affine velocity model We model the image velocity, $\hat{\mathbf{v}}$, as one affine motion for an image region Ω. The number of parameters are then 6 which yields,

$$\hat{\mathbf{v}}(\mathbf{u}, \mathbf{x}) = B(\mathbf{x})\mathbf{u} = \begin{bmatrix} a + bx + cy \\ d + ex + fy \end{bmatrix} \tag{13}$$

$$\mathbf{u} = (a, b, c, d, e, f)^T, \quad B(\mathbf{x}) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x & y \end{bmatrix}$$

where a, b, c, d, e, f are unknown scalar constants.

Solving for the affine image velocity To solve for a velocity field we must use a model for the same affine image velocity, and if we let $\hat{\mathbf{v}}$ be parameterized with $\mathbf{u}$, giving $\hat{\mathbf{v}} = \hat{\mathbf{v}}(\mathbf{u}, \mathbf{x})$, we can pose a weighted least squares minimization problem to solve for $\mathbf{u}$,

$$\min_{\mathbf{u}} = \sum_{\mathbf{x} \in \Omega} w(\mathbf{x}) R(\mathbf{u}, \mathbf{x})^2 \tag{14}$$

where ω is a weight function and Ω is a region of interest in the image where the parameterized velocity model should hold.

Solving for the affine motion parameters The weighting function w in the minimization in (14) is chosen as the gradient magnitude squared, i.e. $w(\mathbf{x}) = \nabla I(\mathbf{x})^2$. Then we have the following minimization problem,

$$\min_{\mathbf{u}} \sum_{\mathbf{x} \in \Omega} (\nabla I(\mathbf{x}) \cdot \hat{\mathbf{v}}(\mathbf{u}, \mathbf{x}) + I_t(\mathbf{x}))^2. \tag{15}$$

It is implicit in this minimization formulation that regions with little/no velocity information contributes less/nothing at all to the computed $\hat{\mathbf{v}}$. By using (13) and (15), a region $\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and measurements of the gradients and time derivatives in the image, we get the following linear equation system,

$$V^T V \mathbf{u} = V^T \mathbf{q} \tag{16}$$

where

$$V = \begin{bmatrix} I_x(\mathbf{x}_1) & I_x(\mathbf{x}_1)x_1 & I_x(\mathbf{x}_1)y_1 & I_y(\mathbf{x}_1) & I_y(\mathbf{x}_1)x_1 & I_y(\mathbf{x}_1)y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_x(\mathbf{x}_n)x_n & I_x(\mathbf{x}_n)y_n & I_y(\mathbf{x}_n) & I_y(\mathbf{x}_n)x_n & I_y(\mathbf{x}_n)y_n \end{bmatrix}$$

$$\text{and} \quad \mathbf{q} = \begin{bmatrix} -I_t(\mathbf{x}_1) \\ \vdots \\ -I_t(\mathbf{x}_n) \end{bmatrix}.$$

(16) is a 6 × 6 symmetric positive semi-definite system, with the 6 elements of $\mathbf{u}$ as unknowns. This is shown explicitly in (17) with all the sums performed over the region to which we want to fit the model. It becomes definite as soon as there are more than one gradient direction present in the region over which the minimization is performed.

$$\begin{bmatrix} \sum I_x^2 & \sum I_x^2 x & \sum I_x^2 y & \sum I_x I_y & \sum I_x I_y x & \sum I_x I_y y \\ & \sum I_x^2 x^2 & \sum I_x^2 xy & \sum I_x I_y x & \sum I_x I_y x^2 & \sum I_x I_y xy \\ & & \sum I_x^2 y^2 & \sum I_x I_y y & \sum I_x I_y xy & \sum I_x I_y y^2 \\ & & & \sum I_y^2 & \sum I_y^2 x & \sum I_y^2 y \\ & symmetric & & & \sum I_y^2 x^2 & \sum I_y^2 xy \\ & & & & & \sum I_y^2 y^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = - \begin{bmatrix} \sum I_t I_x \\ \sum I_t I_x x \\ \sum I_t I_x y \\ \sum I_t I_y \\ \sum I_t I_y x \\ \sum I_t I_y y \end{bmatrix} \tag{17}$$

References

[1] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. In *1st ICCV*, pages 35–54, 1987.

[2] R. Bajcsy. Active perception vs. passive perception. In *3rd IEEE Workshop on Computer Vision*, pages 55–59, 1985.

[3] D. H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[4] P. J. Burt. Attention mechanism for visoin in a dynamic world. In *9th ICPR*, pages 977–987, November 1988.

[5] J. J. Clark and N. J. Ferrier. Modal control of an attentive vision system. In *2nd ICCV*, pages 514–523, 1988.

[6] S. Coren, L. M. Ward, and J. T. Enns. Attention. In *Sensation and perception*, chapter 15. Harcourt Brace College Publishers, 1993.

[7] S. Coren, L. M. Ward, and J. T. Enns. Brightness and spatial frequency. In *Sensation and perception*, chapter 4. Harcourt Brace College Publishers, 1993.

[8] G. W. Humphreys and V. Bruce. Visual attention. In *Visual Cognition*, chapter 5. Lawrence Erlbaum Associates, Publishers, 1989.

[9] M. R. M. Jenkin, A. D. Jepson, and J. K. Tsotsos. Techniques for disparity measurement. *CVGIP: Image Understanding*, 53(1):14–30, January 1991.

[10] A. D. Jepson and M. R. M. Jenkin. The fast computation of disparity from phase differences. In *CVPR*, pages 398–403, June 1989.

[11] D. G. Jones and J. Malik. A computational framework for determining stereo correspondence from a set of linear spatial filters. In *2nd ECCV*, pages 395–410, May 1992.

[12] A. Maki. *Stereo vision in attentive scene analysis*. PhD thesis, Royal Institute of Technology, 1996. ISRN KTH/NA/P–96/07–SE.

[13] A. Maki, P. Nordlund, and J.-O. Eklundh. A computational model of depth-based attention. In *13th ICPR*, volume IV, pages 734–739, 1996.

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

[15] A. Maki, T. Uhlin, and J.-O. Eklundh. Phase-based disparity estimation in binocular tracking. In *8th SCIA*, pages 1145–1152, May 1993.

[16] 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.

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

[18] P. Nordlund and T. Uhlin. Closing the loop: Detection and pursuit of a moving object by a moving observer. *IVC*, 14(4):265–275, 1996.

[19] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE-PAMI*, 15(4):353–363, 1993.

[20] K. Pahlavan and J.-O. Eklundh. A head-eye system - analysis and design. *CVGIP:Image Understanding*, 56(1):41–56, July 1992.

[21] M.I. Posner and S.E. Peterson. The attention system of the human brain. *Annual Review of Neuroscience*, 13:25–42, 1990.

[22] P. A. Sandon. Control of eye and arm movements using active, attentional vision. In K. L. Boyer, L. Stark, and H. Bunke, editors, *Application of AI, Machine Vision and Robotics*, pages 203–223. World Scientific, 1993.

[23] T. D. Sanger. Stereo disparity computation using Gabor filters. *Biological Cybernetics*, 59:405–418, 1988.

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

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

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

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

[28] C.-J. Westelius. *Focus of attention and gaze control for robot vision*. PhD thesis, Department of Electrical Engineering, Linköping University, 1995.

[29] R. Wilson and H. Knutsson. A multiresolution stereopsis algorithm based on the Gabor representation. *Proceedings IEE International Conference Im. Proc. and Applic., Warwick. U.K.*, pages 19–22, 1989.

[30] 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.

[31] R. Zabih and J. Woodhill. Non-parametric local transforms for computing visual correspondence. *3rd ECCV*, pages 151–158, 1994.