International Journal of Computer Vision 17(2), 113–135 (1996)


# Dynamic Fixation and Active Perception

KOUROSH PAHLAVAN, TOMAS UHLIN 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*
kourosh@bion.kth.se
tomas@bion.kth.se
joe@bion.kth.se

*Received March 1993; Revised January 1995*

**Abstract.** Fixation is the link between the physical environment and the visual observer, both of which can be dynamic. That is, dynamic fixation serves the task of preserving a reference point in the world, despite relative motion. In this respect, fixation is dynamical in two senses: in response to voluntary changes of fixation point or attentive cues-gaze shiftings, and in response to the desire to compensate for the retinal slip-gaze holding.

The work presented here, addresses the vergence movement and preservation of binocular fixation during smooth pursuit. This movement is a crucial component of fixation. The two vergence processes, disparity vergence and accommodative vergence, are described; a novel algorithm for robust disparity vergence and an active approach for blur detection and depth from defocus are presented. The main characteristics of the disparity vergence technique are the simplicity of the algorithm, the influence of both left and right images in the course of fixation and the agreement with the fixation model of primates. The major characteristic of the suggested algorithm for blur detection is its active approach which makes it suitable for achieving qualitative and reasonable depth estimations without unrealistic assumptions about the structures in the images.

The paper also covers the integration of the two processes disparity vergence and accommodation vergence which are in turn accomplished by an integration of the disparity and blur stimuli. This integration is accounted for in both static and dynamic experiments.

## 1. Introduction

Dynamic fixation is the physical manifestation of the processes engaged in an active vision system. An active observer moves in his environment and continuously changes visual parameters in order to gather appropriate sensory input for the perceptual task at hand; meanwhile, the environment changes in time. Fixation is the link that connects the dynamic observer to the dynamic environment. It is also the means by which the observer can redirect his attention towards the detected events in the environment that require decisions for action or pure observation.

Fixation is quite often considered as the *binocular* process of bringing a 3D target to the crosspoint of the optical axes of the left and right eyes/cameras. However, fixation is principally a *monocular* process that can be defined by a direction and a distance $(\theta, r)$. In the binocular case, the directions and the distances of the right and the left axes are mapped into a common representation; this common line of sight is usually called the cyclopean axis.

The movement of the fixation point along the cyclopean axis is the vergence movement. This movement together with ballistic saccadic movements, smooth pursuit and other less apparent movements form the

dynamic fixation movements of the primates. In a *monocular* biological fixation process, the rotation of the eyes/cameras perform the task of directing the gaze towards the target of interest. In a completely static scene, the distance to the object of interest is obtained mainly by accommodation.

In the *binocular* case, fixation is performed along the common line of sight by at least two processes, accommodative vergence and disparity vergence. In this way accommodative vergence, with blur as the major stimulus, can take advantage of disparity, while the disparity vergence process, with disparity as the major stimulus, can rely also on blur; resulting in a more robust fixation.

The whole process described has long been known. In this paper, we are going to present a computational implementation of it. Other researchers have studied the fixation problem. Clark and Ferrier (1988) implemented a model of the human oculo-motor system, focusing on the control aspects. Coombs (1992) described a real-time model relying on zero disparities. There is however a basic difference between the approach presented here[1] and the work done earlier.

In order to implement a computational model of e.g. the human oculo-motor system, there should be a device capable of simulating its major characteristics. The only such device presently available is the KTH-head. Other systems, although some of them are, in parts, more advanced than the KTH-head, lack either accommodation facility, separate eye modules, foveal simulation capability (like zooming) or have no neck joints. In other words, they are not designed to simulate a primate oculo-motor system to begin with. Another even more important difference is that the traditional approaches with few exceptions, such as Abbott and Ahuja (1988), have concentrated on movements of the eyes and ignored the other parameters of the system.

The KTH-head allows for synchronous control of eye modules, accommodation, iris control and image magnification. Furthermore, the head gives the possibility to adjust the location of the lens axially so that the optical center can be displaced to different places depending on what the optical parameters are and what the task at hand is[2]. Details about the design of the KTH-head can be found in Pahlavan (1993), Pahlavan and Eklundh (1992) and Pahlavan and Eklundh (1994).

In addition to this, we introduce two novel algorithms/implementations, one for disparity vergence and one for depth from defocus. The first algorithm is based on a very simple observation about the geometry of human vergence movements. It uses the variance of left and right images, and thus obtains a sharpness criterion to discard wrong matches. The second algorithm is an application of Subbarao's algorithm for parallel recovery of depth by changing camera parameters see (Subbarao, 1988). We have extended this model and applied it to changes of the accommodation distance. This is an *active* approach to the problem of depth from defocus, because it manipulates the camera parameters in order to obtain new data about the scene. A related approach, which is also active, is based on changing the iris size, which in our opinion is also an active approach. See e.g. Pentland (1987).

The information acquired by depth from defocus is essential to monocular fixation and important for binocular movements. In this work, we have not addressed saccades that constitute the most important visuo-motoric behavior. There are basic problems in analyzing the attentive-voluntary trigger functions for saccadic movements. However, as far as other involuntary movements are concerned, the system described and the experiments accounted for are equally stable under dynamic monocular as well as binocular fixations.

Work on integrating "focus and stereo" related to ours exists; see e.g. (Krotkov, 1989; Abbott and Ahuja, 1988). However, the work presented here should be seen as neither a continuation of nor a contradiction to the previous approaches. This work is purely concerned with the fixation process and does not cover stereopsis at all. Two other aspects distinguish it from earlier work as well. First, we address cue integration (disparity and blur) and process integration (accommodative vergence and accommodative disparity) separately. Secondly, we do not simply combine the cues by averaging their results, but take their qualitative and quantitative nature into consideration. In addition to this, we ensure that fixation as well as integration holds under dynamic conditions without any artificial constraints or precise calibrations. The algorithms and processes involved in the dynamic scenario, e.g. the tracking and stabilizing algorithm, are not described in this article and require a separate presentation.

The active vision paradigm can be motivated by pure technical and computational advantages. However, it can also be viewed in a much broader context that motivates the kind of activeness that is present in biological systems. We have attempted to address this aspect of the active vision paradigm in Pahlavan et al. (1993)

and will here limit us to technical issues like algorithm design and system function. Still, since the mentality behind the system design has its roots in inspiration from biological vision, the reader should be prepared for some leading discussion of biological implications prior to each algorithm or process description.

A number of experiments are performed to elucidate some crucial aspects of the work. One experiment, is primarily related to the work on depth from defocus and is therefore described in the corresponding section. Another set of experiments demonstrates the proposed disparity detection algorithm with and without contribution from a conventional sharpness criterion. A third set of experiments deal with dynamic situations. During these experiments, the binocularly fixated object is moving rapidly, so that preserving fixation requires a cooperation between vergence and smooth pursuit movements and accommodation.

The smooth pursuit process is used in the dynamic experiments to demonstrate two essential points in an active real-time algorithm design; first, it shows that the fixation algorithm is easily integrated with others such as the smooth pursuit without loosing stability; secondly, it unveils the dynamic capacities of the algorithms and their integration.

The organization of this article is as follows. Initially, our concepts of gaze control are described. Then the fixation geometry of human beings, the stimuli for fixation and a psychophysical experiment corresponding to vergence movements in man are discussed. During this description of disparity vergence, we describe the corresponding geometry of our suggested algorithm. In the same order as above, the *computational* techniques for calculation and integration are described. In the end the experiments on dynamic fixation are demonstrated.

## 2. Gaze Control and the Primary Ocular Processes

The architecture behind the control of the gaze in the KTH-head is called "primary ocular processes". The objective of this architecture is to realize a system where many cues and the corresponding processes running on them are easily integrated. The word "primary" refers to the fact that these processes constitute the innermost control loops with visual feedback. The "ocular processes" refer to processes that possess one of the following characteristics

- processes that use the same stimulus for different tasks.
- processes that use different stimuli for the same task.

The processes that use different stimuli for the same task can be run in parallel and perform better, because of the redundancy in the input. Given a limited number of cues, the processes that use the same stimulus and yield different outputs are necessary for enhancing the capabilities of the system in executing several behaviors.

One advantage of this approach is that the system does not need to average the stimuli for cue integration. Hence, the stimuli do not need to be of the same kind. Another advantage is that each process could be small and simple. However the topology of the processes is crucial to their tasks. The following example could serve to elucidate a chain of connections and cues. Variance can be used as one of many cues to estimate blur that is caused by both defocus and motion. It can also be used as a component in computing temporal and spatial disparities through correlation. The disparities and blur can in turn be used for reliable correspondences, etc. The advantages are twofold: the computations are recycled and the network of connections increases the reliability of the system.

Since primary ocular processes can run in parallel, their outputs, i.e. the potential inputs to an integrating process, could attenuate or amplify each other. That is, if the integration process is aware of in what conditions which stimuli are more reliable, it can weigh the inputs accordingly. It is not easy to resolve the conflict situations beforehand. However, in a simple process with e.g. only two inputs, it is much easier to manage this problem. For example, finding correspondences could involve sharpness (or blur) and disparity as two stimuli. Finding correspondences by disparity is more difficult at near distances and much easier at far distances, the opposite is true for blur. This basic information can form a guide-line for the integrator to deal with potential conflicts between the two cues.

The idea is that by using different stimuli which are due to the same phenomenon, we can achieve better estimates and consequently stable performance. A single cue does not need to yield reliable estimates in its whole range of operation. Accommodative and disparity vergence, which can be driven by blur and disparity, are good examples of such processes. These two vergence processes result in accommodation and vergence; these two might in turn confirm or contest

each other. Another example is estimating orientation of surface patches by the grey-level gradient and texture gradient as different stimuli; it is known that these two stimuli do not always agree.

The traditional gaze control approach is to look at the left and right images and find a pair of disparate points, compute the 3D position, re-fixate and continue the loop. This approach could fail e.g. due to self-occlusions, because the existence of *two* corresponding points are assumed in the inverse kinematics of the fixation mechanism. Especially, in a dynamic condition, when objects appear and disappear, the system will be unstable.

The described process oriented architecture has resulted in a completely different approach in the KTH-head. Even images are considered as competitors in delivering information and in this sense, lack of one image results in total dominance of the other one. This dominance is in turn feasible, because there are monocular cues that can deliver coarse depth information in the absence of accurate binocular cues. The images are stabilized by stabilization processes that not only stabilize images for e.g. disparity detection, but also receive this produced disparity information to synchronize themselves binocularly. They generate stable images so that e.g. accommodation under movement could be possible and at the same time they can use the accommodation to filter out the distance blur from the motion blur[3].

The accommodation and disparity vergence processes in our present implementation run in *parallel*. A stabilization process, although feeding the processes involved in fixation, receives input from them in order to maintain *binocular* fixation. Once the pattern of interest is stabilized, it actually works better if the object moves because of the motion blur in the background; the pattern of interest is easily distinguished in the blurred background.

In an earlier implementation on the KTH-head, two *monocular* stabilization processes stabilized the left and right images of the moving object separately; each accommodation process focused on its own stabilized image and the disparity detecting process brought the two stabilized focused images together. The vergence movement was derived by the cooperation between the two latter processes and was a feedback to the *binocular* stabilization again, see (Pahlavan et al., 1992). The transition from monocular data to binocular and the need for simplification of handling partial or total occlusion, urged us to use the concept of cyclopean vision which is central to the present control strategy.

Our current implementation of the primary ocular processes used in the control of the KTH-head, is using a simple cyclopean representation obtained by superposition and the stabilization process runs on this single representation. This approach is elaborated in the last half of the paper and cyclopean representation is discussed in the next subsection. This work does not make any distinction between optokinesis and smooth pursuit processes mainly because of lack of foveal-peripheral vision.

### *2.1. Common Representation in Binocular Vision*

Binocular vision[4] is redundancy in vision and large field of view. The basic characteristic is the ability to fixate in a large field of view, rather than stereopsis. The biological visual systems do not *crash* when they encounter any degree of occlusion or loose one eye! Even if a visual system uses binocularity with a minimal relationship between the left and right images, a common representation for both images is still needed. This common representation can be as complicated as a full 3D reconstruction or as simple as two superimposed images. The important thing is that the representation contains both 2D and 3D information. Traditionally, the representations contain one or the other. The minimal geometric information needed to establish the gaze direction from this representation is that each point in the representation stands for a direction rather than a well localized point in 3D. Having a common representation, we do not need to let one eye be totally dominant all the time, i.e. the fixation mechanism can choose to refer to the information from the left, right or both images, depending on the circumstances and the degree of contribution of each eye to the common representation. Figure 1 illustrates the effect of dominance on the location of the fictitious eye standing for the common representation in man.

The major drawback of using one eye as a leading eye in binocular vision is that the set of points that are occluded in the dominant eye but are present in the other eye cannot be used as fixation points.

The notion of cyclopean vision is used in different ways. The concept is used after Helmholtz for denoting a subjective eye between the left and right eyes, a so called "binoculus". Julesz (1971) has used it as a representation that reveals new data by integrating the two images. In line with Helmholtz, we regard cyclopean vision, as the kind of vision that contains a

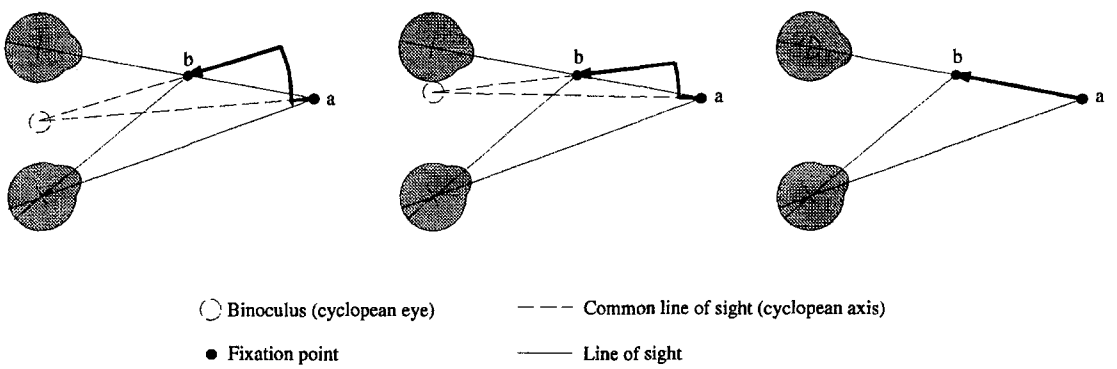


*Figure 1.* The effect of dominance on the location of the imaginary cyclopean eye, the binoculus, in three cases. Left: equal dominance of the eyes. Middle: the left eye is more dominant. Right: the left eye is completely dominant.

common representation of both left and right images and in that sense can be considered as *one* reference image. Following this, we apply an obviously raw but functional approach. An average image of the left and right cameras is the basic input to the system; the symmetries in the images are the cues to vergence and the shift of the symmetries, with respect to the center of the average image, are the cues to version. These issues are discussed in the remainder of the paper.

## 3. Vergence and Version

The process of point to point fixation in primates consists of two separate classes of movements. These movements are vergence (convergence and divergence) and version (conjugate eye movements). In human beings, the two movements are integrated in a complicated manner and not completely independent. However, a strict division into pure version and pure vergence elucidates the geometry behind the mechanism of the two movements. Observations from studying eye movements in human subjects have led to geometric models of different complexity; a simplified illustration of such a model is depicted in figure 2. In this figure, the lines of equal version form a series of rectangular hyperbolae that pass through the sighting center of the two eyes and whose common center is the midpoint of the baseline. The lines of equal vergence are represented by the Vieth-Müller circles passing through the two eyes.

The two movements are associated with two kinds of disparities. Pure version is associated with zero disparity and pure vergence with symmetric disparity; these two qualities are especially interesting for our

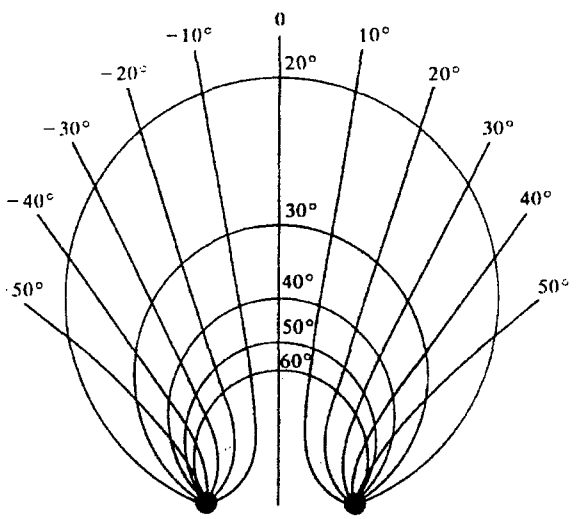


*Figure 2.* The lines of equal vergence (the circles through the eyes) and the lines of equal mean version (the hyperbolae through the eyes). From Carpenter (1988) after Luneburg (1948).

work. The saccadic component of version is a ballistic movement maintaining zero disparity, i.e. occurs ideally along a V-M circle. Contrary to saccades, the vergence movement is image driven and relatively slow due to two different stimuli, blur and symmetric disparity.

### 3.1. *The Geometry of Vergence*

Study of vergence in isolation from version leaves us a simple geometry. This geometry is illustrated in figure 3. Each axis represents the optical axis for the corresponding camera. In practice the central line

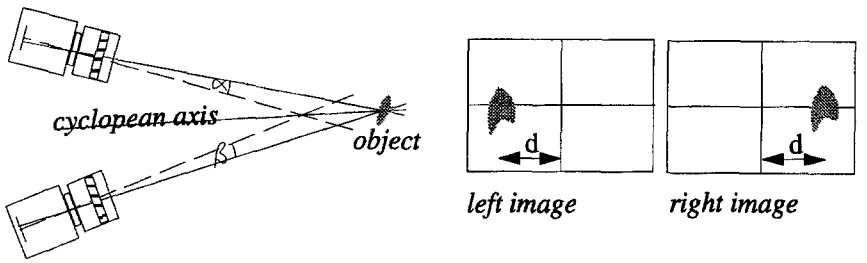


*Figure 3.* The geometry of vergence. Objects located along the cyclopean line are symmetrically positioned in the images with respect to the current fixation point (the center of the images), i.e. $\alpha = \beta$.

representing a common line of sight for the two cameras is the line of sight for an imaginary eye, located somewhere between the two eyes, whose image is the cyclopean representation mentioned earlier. This cyclopean axis, corresponds to one of the hyperbolae depicted earlier in figure 2. The movement of the left and right axes can be represented by the common line of sight, whose direction is a function of the contribution of each image (eye) to the course of movement. In the case when the two cameras are equally dominant, the image of the same object in the scene will be symmetrically located, with the vertical axis through the fixation point as the symmetry axis[5].

The mentioned cyclopean axis plays a central role in the geometry of fixation, and vergence is defined by the displacement of the fixation point along this very axis. The angles $\alpha$ and $\beta$ in figure 3 are equal[6] . The equality of the two angles $\alpha$ and $\beta$ implies that all the objects lying on the cyclopean axis are symmetrically projected into the cyclopean image. That is, the image of the objects that are located along the common line of sight, are found on different sides of the left and right images and the eccentricity of their position in the images is a function of the distance of the projected object from the current point of fixation.

Although we normally do not notice this, we are actually seeing most of the objects in our environment as doublets (diplopia). Despite this fact, we have a very good sense of direction that should also be considered as a part of the cyclopean representation. Besides, there are usually plenty of other attentional stimuli, such as sounds and changes in the scene, that can be sensed as directions rather than positions. Once the direction of interest is known and the saccadic movement is performed, the problem will be how to find the source of interest, i.e. its position; this is when vergence completes the process of fixation.

## 4. Vergence and Its Stimuli

In this section we consider two sources of stimuli for vergence. The vergence mechanism in primates is known to be stimulated by both blur and disparity. Therefore, it is perhaps easier to consider vergence as two different cooperating and competing processes each using a different feedback as stimulus; these two processes are accommodative vergence and disparity vergence. We will describe them separately and discuss their integration as well as how they cooperate and compete.

### *4.1. Accommodative Vergence and Blur*

In 1826, J. Müller showed in a classical experiment, that the vergence movement in human vision is reproducible even in the absence of disparity (Müller, 1826). The procedure is depicted in figure 4. It is obvious from the experiment that changes in blur (even though blur is a monocular stimulus) result in a vergence movement with the perceiving eye as the totally dominant one. Other experiments show that there is a linear relationship between the angle of convergence and the accommodation stimulus, see e.g. (Carpenter, 1988).

Blur can be determined by an analysis of the spatial frequency contents of an image. The estimate is computed as a local response and not a global one, because the degree of blur in an image varies—the objects in the scene are located at different distances. In other words, the computation of the degree of blur must be done locally in a window, and a proper window size is needed. Apart from the optical parameters, the window size depends on the retinal resolution. Currently, the widely utilized CCD cameras have a rather low resolution (compared to human foveal vision) and therefore, they are not capable of detecting small changes in the

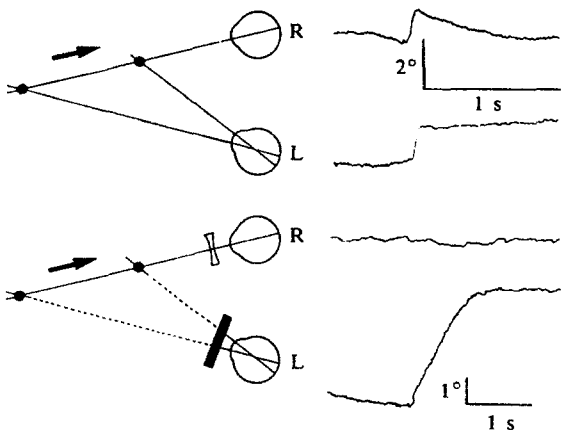


*Figure 4.* Under normal binocular viewing, the vergence/accommodation movement is symmetrical even if the object is moved along the line of sight on one of the eyes (above). Covering one eye (removing disparity) and introducing blur to the eye that has the object along its line of sight, results in a monotonic asymmetric movement of the eye. From Carpenter (1988) after Alpern and Ellen (1956) and Alpern (1957).

frequency domain. Whether focusing and blur detection is feasible or not depends on the relation between local frequency contents and sensor resolution.

Another limitation imposed on accommodation is depth of field which is a function of both sensor resolution and iris size. The smaller the depth of field is, the more precise the depth estimate becomes. In the human case, the very fine resolution of retinal cones in fovea centralis guarantees that at the final steps of accommodation adjustments, an apparently precise accommodation can be achieved.

Static approaches to the problem of depth from blur in most cases use some constraints based on assumptions about what the best focus is; see e.g. (Zhang and Bergholm, 1992) where a step edge is assumed to be a vertically abrupt discontinuity. Such models, however, do not agree with real edges and discontinuities; resulting in difficulties to estimate depth in real images. A good way to go around the problem is to use continuous feedback and keep track of how the sharpness enhancement develops. That is, the focusing should be an iterative procedure rather than a *purely* predictive one. This is applicable both to the process of depth from focus and depth from defocus. Work in our laboratory (Horii, 1992a, 1992b) and experiments by other researchers e.g. (Krotkov, 1989) confirm this procedure. In the psychophysical literature, it is a matter of discussion why the vergence process in our eyes is image-driven and not a ballistic movement like a saccade. Here, we stress the practical significance of the iterative approach.

### 4.2. *Disparity Vergence*

In Section 4.1, we mentioned Müller's experiment that demonstrated how blur, despite being a monocular cue resulted in a vergence movement. However, disparity is a binocular cue and is defined as the angle of correspondence of two associating patterns in the left and the right eye respectively.

The mentioned experiment was evidence for sufficiency of blur as a vergence stimulus. Similar experiments by Fry (1937, 1939), Knoll (1949), Marg and Morgan (1949, 1950)[7] show that the disparity stimulus is a sufficient cue for vergence as well. They also show that it is possible to simulate situations where pure disparity information can suppress the accommodative stimulus and affect the crystalline lens and the pupil size. The procedure is depicted in figure 5 where an insertion of a weak prism in front of one eye, without affecting the accommodation distance, results in a change of fixation.

*We would like to emphasize that experiments like those mentioned here, reveal some aspects in human vision which could be exploited by computational vision for obtaining robust systems. Biological examples here, once more remind us the fact that sufficiency does not always associate with stable performance.*

It was described earlier that lack of knowledge about how sharp a well-focused pattern is, results in a sequential procedure for accommodation; the sharpest pattern is chosen in this strategy. This is also the case with disparity detection. Let us elaborate more on this issue.

Disparity detection is preconditioned upon successful matchings. The problem is that, to our knowledge, there is no *absolute* measure for a good match; especially without presence of distinct features. This is in turn due to the fact that two corresponding patterns in a stereo pair are generally only *qualitatively* similar. The shorter the baseline and the longer the distance to the object, the better is the correspondence.

Although we are only interested in symmetrically located similar patterns in the cyclopean image, there could exist objects in the scene that generate almost similar patterns in the two images as if they were projections from one object positioned along the cyclopean axis (as depicted in figure 6). However, the uncertainty in disparity detection can be decreased by keeping track

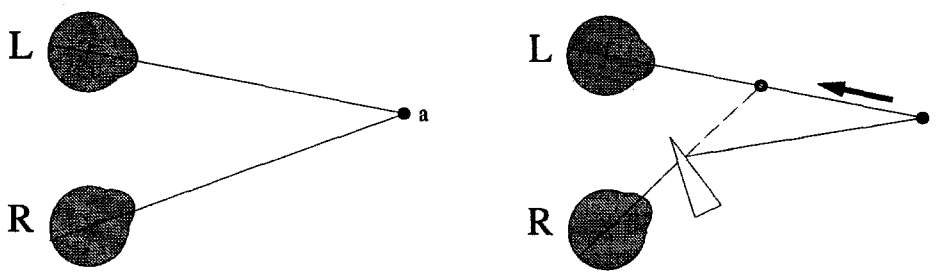


*Figure 5.* Introducing a weak prism in front of one eye when the eyes are fixating a point results in a unilateral vergence to bring the two retinal images together. Redrawn from Carpenter (1988).

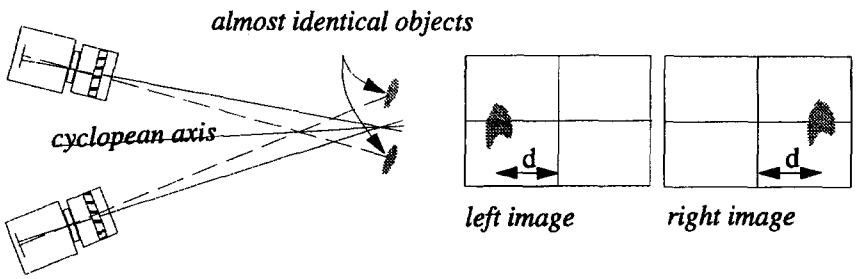


*Figure 6.* In real images, especially when a poor matching algorithm is used, false matches can appear. However, the false peaks do not converge or diverge in the limits they should, under the completion of the vergence/accommodation procedure. The strength of the symmetric approach is that correct matches can only show up at symmetrical locations and can survive the completion of the movement along the symmetry axis.

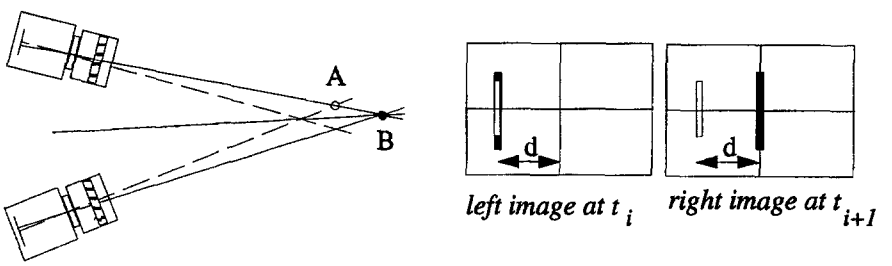


*Figure 7.* A continuous temporal comparison of zero disparities and symmetric disparities associating with the same retinal angle, has the virtue that the vergence process could initiate fusion before vergence is complete. The white object could here be detected by a zero disparity detection of the right and left images at two different temporal states. The dashed lines of sight refer to the first fixation and the solid lines to the next.

of the symmetrically detected matches during the process of convergence or divergence.

The computational procedure above necessitates a slow and image-driven movement. Along with the sequential nature of accommodation, mentioned in Section 4.1, such a phenomenon *might* be the reason why the vergence movement in primates is so slow compared to the fast ballistic saccades.

Our slow sequential model is not providing machine vision with an *absolute* measure for a successful match either. A higher certainty can be obtained by support from other cues like accommodation, as will be discussed in Section 7.3. It should be explicitly stated that the convergence/divergence procedure in the model can be accompanied by successive fusion of appearing zero disparities, i.e. the final matches are not performed on isolated disjunctive points, but rather in the company of successively fusing neighbors. These zero disparity matches have the same uncertainties as the symmetric disparities, but statistically, all these numerous *almost* certain matches result in a much higher certainty. In this context, figure 7 depicts such a probable scenario,

where temporally different zero disparity matches between a pattern observed at time $t_i$ on the fovea of the left image and the pattern at time $t_{i+1}$ on the fovea of the right image are detected.

## 5. Computing Blur

Earlier, we talked about the difference between the active approach and the passive approach in computing blur. An active vision system considers blur as a relative measure, i.e. as a measure for the sharpness degradation or enhancement relative the previous point of fixation.

Blur is naturally measured by the frequency contents over the area at hand. This calls for some kind of frequency analysis. Since Fourier analysis is computationally expensive, one could utilize computational shortcuts or analog approaches that yield the same kind of results from other cues. The discussion in this section begins with criteria of sharpness and continues with finding the point of best sharpness. In the end of the section a computational model for detecting blur and depth from defocus will be suggested.

In the following we speak about *sharpness criterion* and *blur measure* as two different things. The reason is that we want to distinguish between focusing and blur detection. The former refers to maximizing the sharpness e.g. when accommodating, while the latter refers to computing the extension of the blur circle for a specific purpose, e.g. measuring depth from blur without accommodation.

### 5.1. *Sharpness Criteria*

Sharpness is a relative measure. A criterion is needed to measure the change in the sharpness during focusing. Several criteria for sharpness have been implemented and examined in our laboratory by Horii (1992a). In earlier research by Tenenbaum (1970), Jarvis (1976), Krotkov (1989) and others the following have been applied:

***The Tenengrad Criterion.*** This criterion was proposed by Tenenbaum (1970). It estimates the gradient at each image point. The criterion for sharpness is the sum of the gradient magnitudes. The best focus is obtained when the sum attains its maximum. The images are assumed to be normalized. The criterion function is:

$$\sum_x \sum_y S(x, y)^2$$

where $S(x, y)$ is the gradient magnitude in $(x, y)$.

***The Gray-Level Variance Criterion.*** Since different degrees of sharpness of the image discontinuities result in a large variation in the intensity levels, it is reasonable to use the grey-level variance as a sharpness criterion. The formulation of variance in this case could look like:

$$\sigma^2 = \frac{1}{N^2} \sum_{x=1}^{N} \sum_{y=1}^{N} (I(x, y) - \mu)^2$$

where $\mu$ is the mean of the gray-level distribution. The maximum acquired value of $\sigma^2$ is then the criterion for good sharpness.

***The Sum-Modulus-Difference (SMD) Criterion.*** This measure (Jarvis, 1976) is based on the sum of the difference between the neighboring pixels along a scan line. The measure is, in practice, utilized on both horizontal and vertical axes. The SMD-value on each axis can be computed by:

$$\mathrm{SMD}_x = \sum_x \sum_y |I(x, y) - I(x, y - 1)|$$

$$\mathrm{SMD}_y = \sum_x \sum_y |I(x, y) - I(x + 1, y)|$$

and best sharpness is achieved by maximizing the sharpness criterion:

$$\mathrm{SMD} = \mathrm{SMD}_x + \mathrm{SMD}_y$$

### 5.2. *Blur Detection*

The methods presented in Section 5.1 are used to find the best sharpness in the region of interest. This region corresponds to one and the same part of the scene. The idea is to find the *maximum* sharpness on the same pattern and computing depth after completed accommodation. However, in active vision it is often desirable to have a measure of blur rather than a measure of sharpness. Blur detection is traditionally concerned with how fuzzy the image is compared to the *normal* case; this "normal" should then be defined somehow.

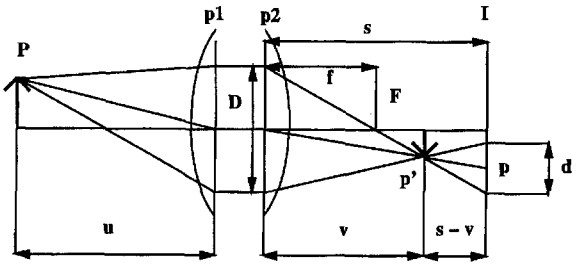


*Figure 8.* The camera geometry and camera parameters; $p1$: first principal plane, $p2$: second principal plane, $P$: object point, $p$: image point, $I$: image detector plane, $s$, $f$, $D$: camera parameters ($s$: distance between the image detector plane and the second principal plane, $f$: focal length of the lens, $D$: numerical aperture of the lens), $u$: the object distance, $v$: screen plane distance, $d$: blur circle diameter.

Here we attempt to use the blur measure for a qualitative depth estimation without complete accommodation and without having any assumptions about the structure of the pattern.

We cannot use accommodation because we normally use blur detection when we have already fixated/accommodated on an object and the blur measure is used to investigate the environment in the periphery of the fixated object.

In the following formulation of the problem, we use Subbarao's development of the problem (Subbarao, 1988). He discussed active blur detection and developed an algorithm that allows for depth recovery through manipulation of camera parameters. Subbarao's problem is not identical to ours. Nevertheless, we could apply the idea to the particular case of manipulating the image distance. We introduce an approach to avoid the heavy computations and problems with variations in texture. The significant point is that robust qualitative range information can be achieved by an active approach and we claim that this kind of information is useful.

The effect of defocusing can be described by a point spread function. Let $P$ be a point in the scene and $p$ be its focused image (see figure 8). If $P$ is not in focus then its refracted image becomes a circular image called the *blur circle*. A point-spread function can represent the structure of this circle.

The relation between the position of $P$ and $p$ is given by the lens formula:

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v} \tag{1}$$

in the formula, $f$ is the focal length, $u$ is the distance from the first principal plane to the object and $v$ is the distance from the second principal plane to the sharp image. The diameter of the blur circle $d$, can be derived from the lens formula and the geometrical relationships depicted in figure 8:

$$d = Ds\left(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}\right)$$

where $D$ is the diameter of the lens and $s$ is the distance from the second principal plane to the image plane. According to projective optics, the intensity distribution within the circle is almost constant.

The pill box point spread function is the ideal model. However, due to diffraction, the variation of the wavelength of light in the image, aberrations and other effects, Pentland (1987), Subbarao (1988) and many other researchers have advocated the use of the following Gaussian function as the point-spread function.

$$h = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{x^2+y^2}{\sigma^2}}$$

The parameter $\sigma$ is the spread parameter whose value depends on the intrinsic parameters of the imaging system. This dependence is embodied in the calibration constant $k$:

$$\sigma = kd \quad \text{or} \quad \sigma = kDs\left(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}\right) \tag{2}$$

Let $g(x, y)$ be the observed image of an object on the sensor, and $f(x, y)$ be the corresponding sharp image. Then $g$ can be formulated as the convolution of $f$ and the Gaussian kernel $h$.

$$g(x, y) = h(x, y) \otimes f(x, y)$$

The power spectral density for a Gaussian point spread

function is

$$P(\omega, \upsilon) = GG^* = HH^*FF^* = e^{-(\omega^2+\upsilon^2)\sigma^2}FF^* \tag{3}$$

where $F$, $G$ and $H$ are the Fourier transforms of $f$, $g$ and $h$. That is, the exponential decay of the power spectrum can be used as a measure of blur. By comparing the spectral density between sequences of images distinguished by the manipulated parameters $s$, $f$ and $D$, it is possible to compute depth. In our particular case, we are interested in the variation of blur due to a small change in the accommodation distance (manipulating $s$).

We need two images with which we can compute the variations in the blur parameter, i.e. $d\sigma$ and through it the corresponding depth. Consider two successive images with camera settings $s$ and $s+ds$. From Eq. (2) we have

$$d\sigma = kD\left(\frac{1}{f} - \frac{1}{u}\right)ds \tag{4}$$

Variations in the spectral density can be studied by differentiating Eq. (3)

$$dP = -2(\omega^2 + \upsilon^2)P\sigma d\sigma \tag{5}$$

which can be rearranged to

$$\sigma d\sigma = -\frac{1}{2}\frac{1}{\omega^2 + \upsilon^2}\frac{dP}{P} \tag{6}$$

Let

$$C = -\frac{1}{2}\frac{1}{\omega^2 + \upsilon^2}\frac{dP}{P} \tag{7}$$

Now, we can use the association between $\sigma$ and camera parameters through Eqs. (2), (4), (6) and (7)

$$\sigma kD\left(\frac{1}{f} - \frac{1}{u}\right)ds = k^2D^2\left(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}\right)\left(\frac{1}{f} - \frac{1}{u}\right)ds = C$$

$k^2D^2ds$ can be combined into a single system dependent value $K$. Here we conclude the application of Subbarao's method to our particular case by replacing $X = \frac{1}{f} - \frac{1}{u}$ and solving the equation above for it. Since $f < u < \infty$ in normal experimental settings, the solution will be unique.

$$X = \frac{1}{f} - \frac{1}{u} = \frac{K + \sqrt{K^2 + 4KCs^2}}{2Ks} \tag{8}$$

Thus, the distance to the object is:

$$u = \frac{1}{\frac{1}{f} - \frac{K+\sqrt{K^2+4KCs^2}}{2Ks}}$$

At this stage, we are confronted with two major computational problems: computing the spectral densities and computing $C$. Fourier transforms are computationally expensive and in our particular case unnecessary, because we only need the spectral density at our points of interest rather than the transformations. As an approximate solution, we propose an adapted version of a method suggested by Pentland et al. (1989).

The next problem is computing $C$. The problem here is that due to the mentioned approximation, $C$ is a function of the texture; in a complicated scene, it is necessary for the algorithm to be invariant to the structure of the texture, although the presence of texture is a must. We can solve these two problems as follows.

$f_s$ and $f_{s+\Delta s}$ are two images distinguished by a small change of $s$. These images are band pass filtered by a Laplacian kernel and the square of the pixel values are summed up to create the Laplacian powers $P_s$ and $P_{s+\Delta s}$.

$$P_s = \sum_x \sum_y (f_s \otimes L)^2$$

$$P_{s+\Delta s} = \sum_x \sum_y (f_{s+\Delta s} \otimes L)^2$$

The ratio $dP/P$ can then be calculated as

$$\frac{dP}{P} = \frac{2(P_{s+\Delta s} - P_s)}{P_s + P_{s+\Delta s}}$$

The parameter $C$ in Eq. (7), can be evaluated by calibrating $k_f = \omega^2 + \upsilon^2$. However, as described before, $k_f$ depends on the local texture. We use a *dynamic technique* to calibrate this value by blurring the original image through convolution by a Gaussian kernel $G$; as if the image is defocused in a third step.

$$f_g = f_s \otimes G$$

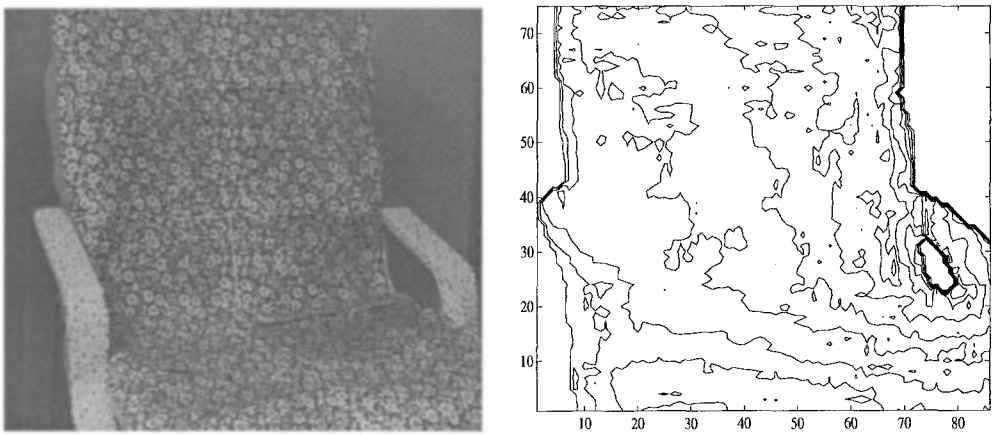


*Figure 9.* The image of a chair and the level curve representation of the result obtained by active approach to depth from defocus. The chair is at a distance of about 2 meters. Two frames are used here, and the difference between the two images is a small change in the accommodation distance. Each level curve stands for a relative distance of 100 mm.

The Laplacian power of $f_g$ is calculated as

$$P_g = \sum_x \sum_y (f_g \otimes L)^2$$

The relation between the Laplacian power of the original image and the synthetically blurred image, $dP_g/P_g$, is calculated as above

$$\frac{dP_g}{P_g} = \frac{2(P_g - P_s)}{P_s + P_g}$$

Since the blurring was a constant Gaussian operation, Eq. (7) will give us

$$k_{\text{const}} = -\frac{1}{2}\frac{1}{k_f}\frac{dPg}{Pg}$$

Thus, the value of $k_f$ is calibrated to

$$k_f = -\frac{1}{2}\frac{1}{k_{\text{const}}}\frac{dP_g}{P_g}$$

The value of the $k_{\text{const}}$ can be included in the constant $K$ in Eq. (8) so that $K$ will be re-valuated to

$$K = k^2 D^2 ds\, k_{\text{const}}$$

This $K$ is the only parameter that require an a priori calibration. In the same way, $C$ in Eq. (7) is reformulated to

$$C = -\frac{1}{2}\frac{1}{k_f}\frac{dP}{P} = \frac{P_g}{dP_g}\frac{dP}{P} \qquad (9)$$

This way the distance $u$ is extracted from the blur information, quickly and robustly. It should, however, be underlined that the depth information obtained from blur and sharpness are qualitative; this is especially true in the case of blur. To complete the discussion about depth from blur, the experiment illustrated in figure 9 is worth referring to. The approach above is applied to a sequence of two images, one of which is depicted in the figure. The other image is identical to this one except for a slight change in the accommodation distance. The relative change in the blur is detected in the two images and the level curve depth map in the figure is generated. As shown in the level curve image, the algorithm has some problem around the contours of the chair. This is due to two factors: the magnification effect of the focusing mechanism in the conventional lenses and the processing window. The former problem is solved on the KTH-head by allowing a small change of the focal length. However, as a future extension to the work, it is more appropriate to address the problem together with simultaneous iris manipulations as well.

## 6. Computing Disparity

Matching can be performed in different ways and on different kinds of features. Nevertheless the fundamental problem is to find similarity of patterns. In our experiments, we have used a normalized correlation method for finding the best match between the two images. Later in Section 9.3, we will also briefly discuss a phase based disparity detection method.

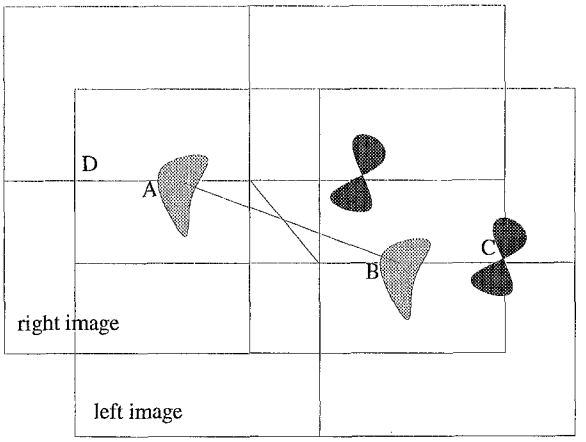


*Figure 10.* The symmetric disparity search on the left and right images. Each position in an image is correlated to the associating symmetric position in the other image. Here the correlation between position A in the right image with position B on the other side of the center of the image in the left image would yield a high peak. The same is not valid for points C and D, suggesting that the pattern at A and B are the images of the same object located on the common line of sight. The following convergence will keep track of the peaks and confirm the correlation results both by checking the changes in the sharpness by the variance method discussed earlier and by checking if the tracking succeeds, i.e. if the two patterns come together at the center of the images.

Figure 10 demonstrates how the normalized correlation is performed. For each region on the left side of an image, there is another region on the right side of the other image that could contain the right pattern. If the correlation between these two regions yields a high peak, the match is accepted. Normalized correlation is given by:

$$C_{xy} = \frac{\text{Cov}[L, R]}{\sqrt{\text{Var}[L]\text{Var}[R]}}$$

where Cov$[L, R]$ is the covariance and Var$[L]$ is the variance.

Later in Section 9, the variance is used as a criterion for sharpness. The reason is simply that the variance is already computed for disparity detection and therefore can be used directly for computing the degree of sharpness in the window. This is an expected integration characteristic of primary ocular processes that allows the processes to share computational steps.

## 7. Integrating Accommodation and Disparity

In the end of the last century there was a belief that accommodative vergence was the controlling component of vergence and the disparity vergence had only a supplementary role in the process (as in Maddox, 1886). However, later experiments (like the one mentioned in Section 4.2 and illustrated in figure 5) show that the relation is more complex, as R.H.S. Carpenter puts it:

> ... it is clear that the accommodation convergence component in fixating near objects is substantial, accounting for rather more than half the total response .... However, it is in any case misleading to think of vergence as being simply the linear sum of two independent stimuli, because of the complex feedback relationships that exist between vergence and accommodation on the one hand, and disparity and blur on the other. These relationships appear to be perfectly symmetrical, in that changes in disparity elicit not only vergence but also accommodation ... (Carpenter, 1988, p. 109).

It is not explicitly motivated here why there should be such a cue integration between blur and disparity. However, it should be noted that the cue integration has in fact a stabilizing effect on the process of fixation. Figure 11 illustrates how the two blur and disparity stimuli can be integrated to realize a mutual effect on both accommodation and vergence. The figure does not clarify how the error from each stimulus affects the whole system. As long as the two stimuli agree, i.e. the two stimuli increase or decrease together, small amounts of noise do not matter, because the process is a closed loop and as such, it can overcome small amounts of noise and enter a steady state.

However, the two stimuli result in two sets of actions that can be seen as two different processes—accommodation and vergence. There are situations where these two processes come into a conflict situation. Note that we are now talking about the situations where the *processes* are in conflict and not the stimuli. We have already mentioned the classical experiments where researchers have isolated one stimulus and analyzed its effect on accommodation and vergence.

The relationship between accommodation and vergence can be divided into four categories:

- The two processes agree and both of them are wrong. The range for best accommodation is too large and the disparity detection fails inside this region.

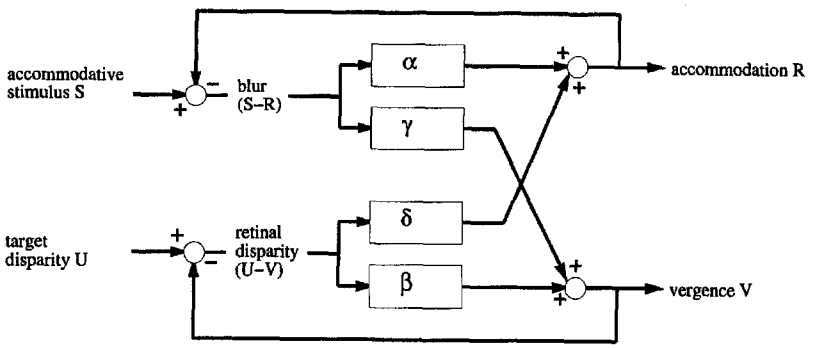


*Figure 11.* A model of the relationship between accommodative stimulus ($S$), target disparity ($U$), accommodation ($R$) and vergence ($V$): $R = \alpha(S - R) + \delta(U - V)$; $V = \beta(U - V) + \gamma(S - R)$. From Carpenter (1988).

- The two processes agree and both of them are right.
- The two processes disagree because the accommodation process is conflicting. Illusions, effort of will and strong (although wrong) disparity stimulus make vergence rely on disparity.
- The two processes disagree because there are many candidates for a match.

One would wonder how two processes that use the same stimuli can disagree. The answer is that they would not, if the disparity and blur stimuli were unique, i.e. if there were not other powerful candidates competing with them (e.g. in the case of repeated patterns, specular and transparent surfaces and alike). Later in this section, we will discuss two cases where the stimuli are misleading the processes.

If we only have these two processes to integrate[8], then the first category would lead to a system reset or binocular rivalry, and nothing could be done about it. Total darkness is an example of the first case when the eyes move to the resting position (focusing at infinity); exposing the eyes to different images is an example of the other case (binocular rivalry).

In the second case, everything is in order: since the system is a closed loop, it can manage to filter out the small but inevitable noise in the stimuli.

The third and the fourth cases are the interesting ones and are discussed in the following subsections.

### 7.1. *Erroneous Detection of Disparity*

An erroneous disparity detection could have two reasons—repeated patterns and improper matches. An example of the first case is looking at wall paper or a fence from a near distance so that most of the retina is covered by the repeated pattern[9]. Such an example is addressed in Pahlavan et al. (1992). In this case the accommodation process overrides the vergence process.

An example of the second case is when no corresponding matches are found. This is the case when the target is occluded in one of the cameras. Implicitly, this means that there already exists an interesting pattern or feature that is chosen in the image of the other eye. That is the seeing eye becomes totally dominant and the disparity stimulus does not contribute to the accommodation process. The blur stimulus becomes dominant in the cue integration and the accommodation process overrides vergence. A brief description of the two cases is:

***Repeated Pattern.*** The disparity stimulus is not unique due to the existence of repeated patterns → accommodation rejects false disparity candidates outside the depth of field.

***No Proper Match.*** The dominance is decided by knowing in which eye the pattern of interest is chosen → accommodation in that eye overrides both the accommodation in the occluded eye and the vergence.

### 7.2. *Conflict in Accommodation*

The conflict in the estimation of accommodation distance can be categorized into two cases. Since accommodation is a monocular process and a well-posed problem, it is perhaps not correct to talk about erroneous accommodation. The error is not principally in the accommodation process itself. It is more proper to talk about the confusion arising from the internal conflict in the *binocular* accommodation, i.e. the conflict between the right and the left eyes' accommodations. A typical example of this is when one eye is occluded.

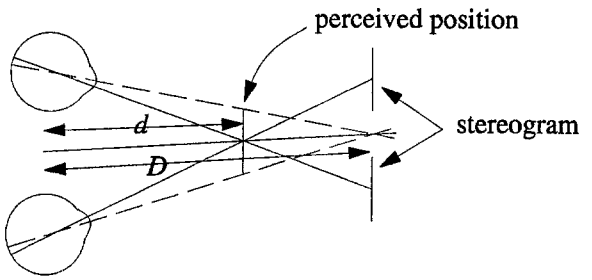


*Figure 12.* Looking at a stereogram by crossing the eyes, gives the impression that the fused imaginary object is nearer to the subject, compared to the real location of the stereogram. Although the accommodation distance is perceived correctly, the distance due to disparity overrides accommodation.

As the case was with the disparity error due to occlusion, the eye which is not occluded should become totally dominant, unless the occluding object, or other objects free from the occlusion are chosen for fixation. In the latter case, a saccadic movement is also involved, but we will not discuss that case further.

The other case is when the common accommodation is correct, but the binocular accommodation distance does not agree with vergence. The situation here is rather delicate and therefore needs a bit more elaboration based on the example of a stereogram. Looking at stereograms by crossing the eyes, puts accommodation and vergence in conflict (see figure 12). It should be noted here that, for human beings, fusing stereogram images does not affect the sharpness of the images, i.e. the accommodation process per se is intact and functional; it is the perceived depth which creates the conflict. Moreover, the accommodation process receives erroneous disparity stimulus which is also in conflict with the blur stimulus. Nevertheless, it is obvious that one can fuse the images. An important factor here is that the disparity cue is unique and very stable in all stereograms[10]. Furthermore, it should be noted that without special glasses, fusion is unstable and can only be preserved voluntarily. A brief classification of the two cases can be given as:

***Occlusion of One Eye.*** No disparity stimulus to accommodation or vergence → the seeing eye is totally dominant using only blur stimulus.

***Binocular Suppression of Focusing.*** Strong and unique disparity stimulus is present → accommodative sense of depth is suppressed (sharpness is unchanged). The situation will be unstable as soon as a new fixation is performed.

### *7.3. Integration*

Now that we have described the major cases of conflict between accommodation and vergence, we can integrate the two processes so that the outcome is robust. The following process integrates the two processes and decides which one should override the other and how:

```
loop forever;
   d ← a quantitative value for depth from vergence;
   D ← a spatial region suggested by
         binocular accommodation;
   if d unique then
      if d ∈ D then
            perceived depth = d;
      else
            if d stable then
               perceived depth = d;
            else
               binocular rivalry
   else
      if D_left ~ D_right then
            perceived depth = d_computed in depth of field.;
      else
            binocular rivalry;
end loop.
```

In practice, as will be demonstrated in Section 9, the procedure is not directly applicable as it is described above. One would rather let blur and disparity be computed in several steps towards the assumed fixation point. The agreement of the processes for computation of blur and disparity is then checked in several steps. The movement of the symmetric disparities that are confirmed by accommodation should persist until their convergence towards the center of the cyclopean representation or the divergence from it is established.

## 8. Computing Depth of Field

The dependency of disparity on blur can be summarized by saying that the disparity detection can be limited to the area in the image associated to the depth of field of the imaging system. Therefore, it is very important to know where this area for matching in the image is. In the small area defined by the depth of field, it is the disparity alone which is responsible for finding the correct correspondences.

Figure 13 illustrates the geometry of the imaging system discussed in the remainder of this section. From

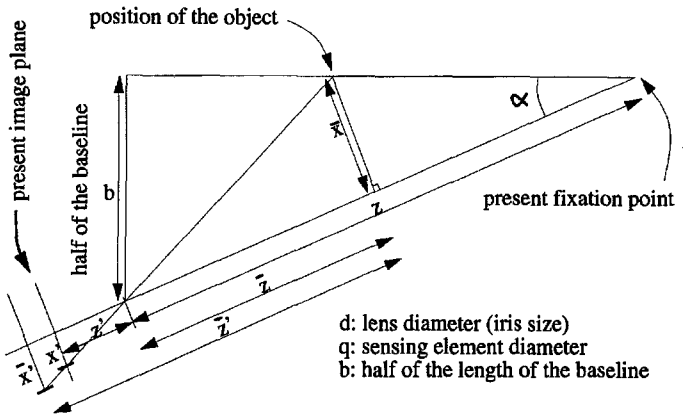


*Figure 13.* The geometry behind the computation of depth of field and its relation to the area in the image where the disparity is the only cue.

the figure we have:

$$\tan\alpha = \frac{\bar{x}}{z-\bar{z}} \qquad \sin\alpha = \frac{b}{z}$$

If $\alpha$ is small (which means that the paraxial approximation is also valid):

$$\bar{x} = b\frac{z-\bar{z}}{z} \tag{10}$$

Also

$$\frac{\bar{x}}{\bar{z}} = \frac{\bar{x}'}{z'} \tag{11}$$

Equations (10) and (11) together yield:

$$\bar{x}' = bz'\left(\frac{1}{\bar{z}} - \frac{1}{z}\right) \tag{12}$$

According to Horn (1986), the depth of field can be defined as:

$$\frac{d}{z'}|\bar{z}' - z'| < q$$
$$d\left|\frac{\bar{z}'}{z'} - 1\right| < q \tag{13}$$

and according to the lens formula, we also have

$$\frac{1}{z'} + \frac{1}{z} = \frac{1}{f} \quad \text{and} \quad \frac{1}{\bar{z}'} + \frac{1}{\bar{z}} = \frac{1}{f}$$

These two equations together yield:

$$\frac{\bar{z}'}{z'} = \frac{\frac{1}{f} - \frac{1}{z}}{\frac{1}{f} - \frac{1}{\bar{z}}}$$

Applying this result to Eq. (13) gives:

$$d\left|\frac{1}{f} - \frac{1}{z} - \left(\frac{1}{f} - \frac{1}{\bar{z}}\right)\right| < q\left(\frac{1}{f} - \frac{1}{\bar{z}}\right)$$
$$d\left|\frac{1}{\bar{z}} - \frac{1}{z}\right| < q\left(\frac{1}{f} - \frac{1}{\bar{z}}\right)$$

For $z > \bar{z}$ we will have:

$$(d+q)\frac{1}{\bar{z}} < q\frac{1}{f} + d\frac{1}{z}$$
$$(d+q)\left(\frac{1}{\bar{z}} - \frac{1}{z}\right) < q\left(\frac{1}{f} - \frac{1}{z}\right) = q\frac{1}{z'}$$

In Eq. (12)

$$\bar{x}' < b\frac{q}{d+q}$$

and for $z < \bar{z}$ equivalently. Finally considering $d \gg q$

$$|\bar{x}'| < \frac{bq}{d}$$

That is, the area in which disparity is the only cue to a correct match, is restricted to $bq/d$ and therefore, it is only a function of the length of the baseline, the iris opening and the sensor resolution.

## 9. Experiments

To illustrate the principles described above, we demonstrate the results from two sets of experiments on static targets. The static scene allows the reader to follow the sequence of images. In the first experiment, the

disparity patterns are clear and the object is chosen so that the blur contribution is not directly decisive. In the second experiment, the patterns are much more complicated and natural; there is a major risk for erroneous disparity detection.

In Section 9.3, the dynamic case is demonstrated, where fixation is performed in real-time (25 Hz) and two stabilized image sequences are shown without vergence (to see the symmetries) and with vergence movement added; each sequence with different disparity detection methods.

### 9.1. *An Experiment with Clear Disparity Patterns*

We simulated the divergence movement from a point close to the head (104 cm) to a point on an object further away (185 cm). During the divergence movement best focus was kept on the point of fixation. Left and right images were acquired five times during this movement uniformly distributed in the range. Between each image pair normalized correlation was performed as a matching criterion. This was done in the symmetric manner described earlier and the highest peak was extracted. A sharpness criterion is associated to this correlation peak in each image; the sharpness criterion is the grey-level variance. It should be higher when an object is better focused. Tracking a potentially correct peak can be performed by applying the following rules in analyzing each image pair:

- A correct peak should continuously and consistently move in the correlation image; towards the center in case the object is in front of the current point of fixation and towards the periphery otherwise.
- If the peak is moving towards the center, the focus values associated with the peak should both increase (decrease if moving outwards).

These rules are both qualitative and do not assume any knowledge about vergence angles or the length of the baseline. However, if we know both the vergence angles and the length of the baseline, we can also predict how the position, the peak and the sharpness would change. This proved to be unnecessary in this case.

Figure 14 shows the result of the experiment. As seen in the figure, the first rule is violated between the first two image pairs. The last three pairs are however consistent from this point of view. Between the third and fourth image pairs, the peak moves correctly towards the center. However, the sharpness criteria

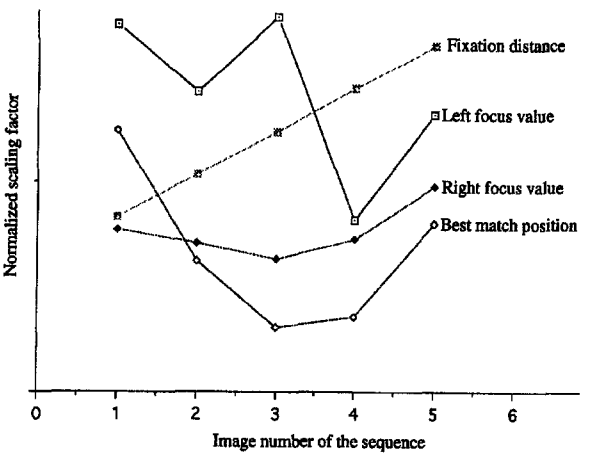


*Figure 14.* The peak position is the position of the maximum correlation peak. A negative value on the peak position is associated with an object beyond the fixation point (positive in front). The value is half of the disparity.

are inconsistent and contradictory; the second criterion is violated. Finally, the changes are consistent between the two last image pairs and the conclusion is then that this is a correctly detected peak and that the vergence procedure has succeeded. It is seen from figure 15 that this is indeed a correct conclusion.

### 9.2. *An Experiment with Direct Involvement of Blur Information*

In the previous experiment the maximum peak along a horizontal line was selected as a correct match. This search was performed without taking sharpness into account. The sharpness can however restrict the search limits drastically, using the following observation. A point in the left image moving away from the center should become more defocused and a point moving towards the center should become sharper (this should hold also for the right image except with the opposite relationship).

This restriction is used to mask off the regions in the correlation image where the rule above is violated. The masking procedure is depicted in figure 16. The search is limited to the areas that are not masked. The masking is not applicable in the region defined by the depth of field.

The first correlation image has no part masked since it is the result of the first image pair and therefore a change in the sharpness cannot be computed. In the

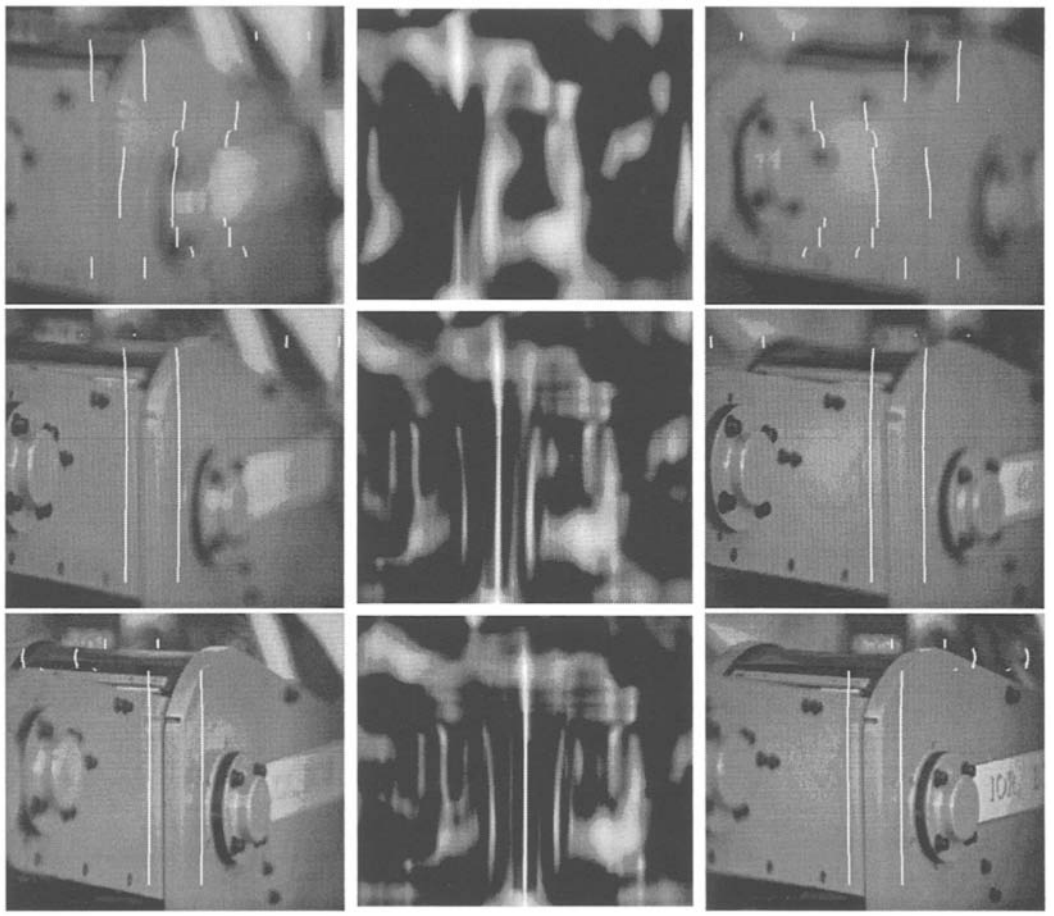

*Figure 15.* The left and right images from the last three divergence steps are shown together with their correlation image. The white lines in the left and right images are indicators for how successful matchings are and stand for the positions corresponding to the highest peak in the correlation image. The correlated areas in the images are depicted by the two parallel lines; the area inside the white lines is the correlation area. As the divergence movement proceeds, the portions of the white lines are gathered around the correct fixation area.

following images, the previous mask is superimposed in order to incorporate earlier focus violations. In all but the last image the mask is only effective in the right half of the correlation image, because the object is getting better focused. In the last image pair the best focus position is passed and the left half is also masked off.

Since the system has no previous observation about the new data that appear in the lateral boundaries of the images, these areas cannot be masked off. The non-masked band in the leftmost part of the correlation image in the bottom, is due to this phenomenon.

The sequence in figure 16 begins from the top and is completed at the time when the third pair of images (from top) are grabbed. The next step of divergence (the image pair at the bottom) is the overshoot that confirms the validity of the fixation at the third image pair by detecting that the sharpness is decreased in the last image pair. This sequence demonstrates also the case when the disparity detection fails in the region defined by depth of field. The region left in the masked region in the left part of the correlation image in the bottom is caused by the sharp structure image due to the appearance of the object *behind* the leaves that suddenly pop up in the left image when the overshoot is realized. Fixating at this background structure can be executed only by the left camera accommodation.

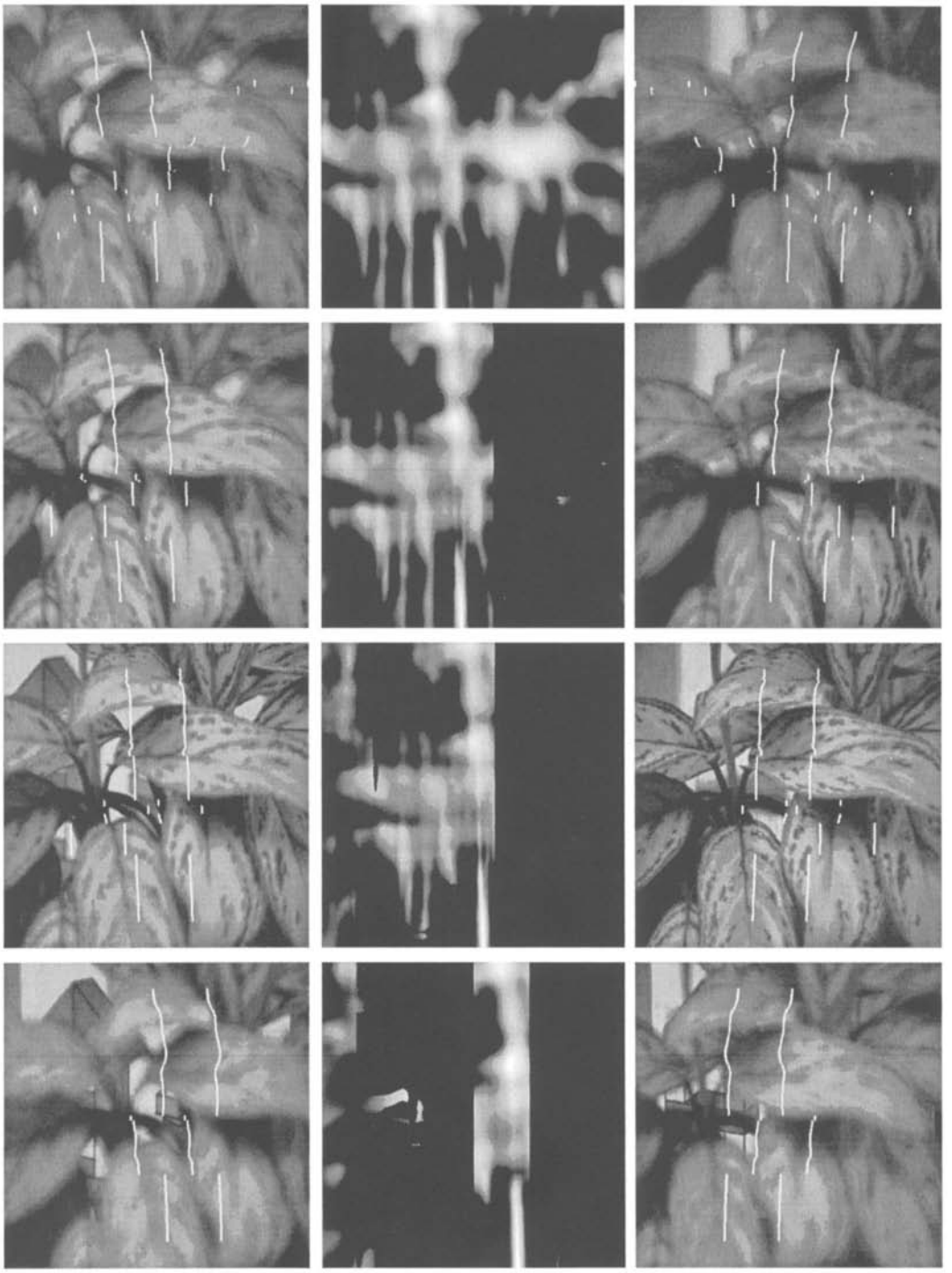

*Figure 16.* The left, right and corresponding correlation images from the last four divergence steps are shown in a sequence from the top. The variance method is used here as a criterion for sharpness detection.

*Figure 17.* A sequence of dynamic tracking of a moving person in real-time. The movement is at a speed of about 50 degrees per second. Every fifth frame is depicted from top-left to bottom-right. In order to see the symmetries, the fusing vergence movements are disabled.

### 9.3. *Two Dynamic Experiments: Verging on Moving Objects*

In the vergence experiments shown previously the environment was static. In order to perform the same experiment in a dynamic environment (i.e. fixate on a moving object) it is essential to maintain the symmetry during the vergence movement. That is done by a *parallel* stabilizing process which compensates for asymmetric movements of the object with lateral saccades or pursuit movements. This can be seen in figure 17, where the symmetrical location of the face/neck pattern in the average image is maintained while the person is moving in front of the camera.

The pursuing process is now totally *independent* of the vergence process, because the symmetry is not affected by the symmetric vergence movements. This shows that the vergence strategy described in this article is functional, not only in a static but also in a dynamic environment. In fact, it seems as if it performs even better dynamically. This is because only the symmetry in the location of the tracked object is maintained in time, while false symmetries from the background, that even survive motion blur, will only occur spuriously.

In order to make the symmetries visible, the motoric contribution of the vergence movements to the tracking sequence depicted in figure 17 has been removed from the control loop.

In another set of dynamic experiments, we have added the vergence movement to the smooth pursuit. Once a target has been successfully verged on and the tracking process has locked onto this target we can expect slow and small changes in the disparity in the center of the image, where the target is. Consequently the small change between frames in time makes it feasible to invoke a much simpler vergence algorithm in order to *keep* the disparity small.

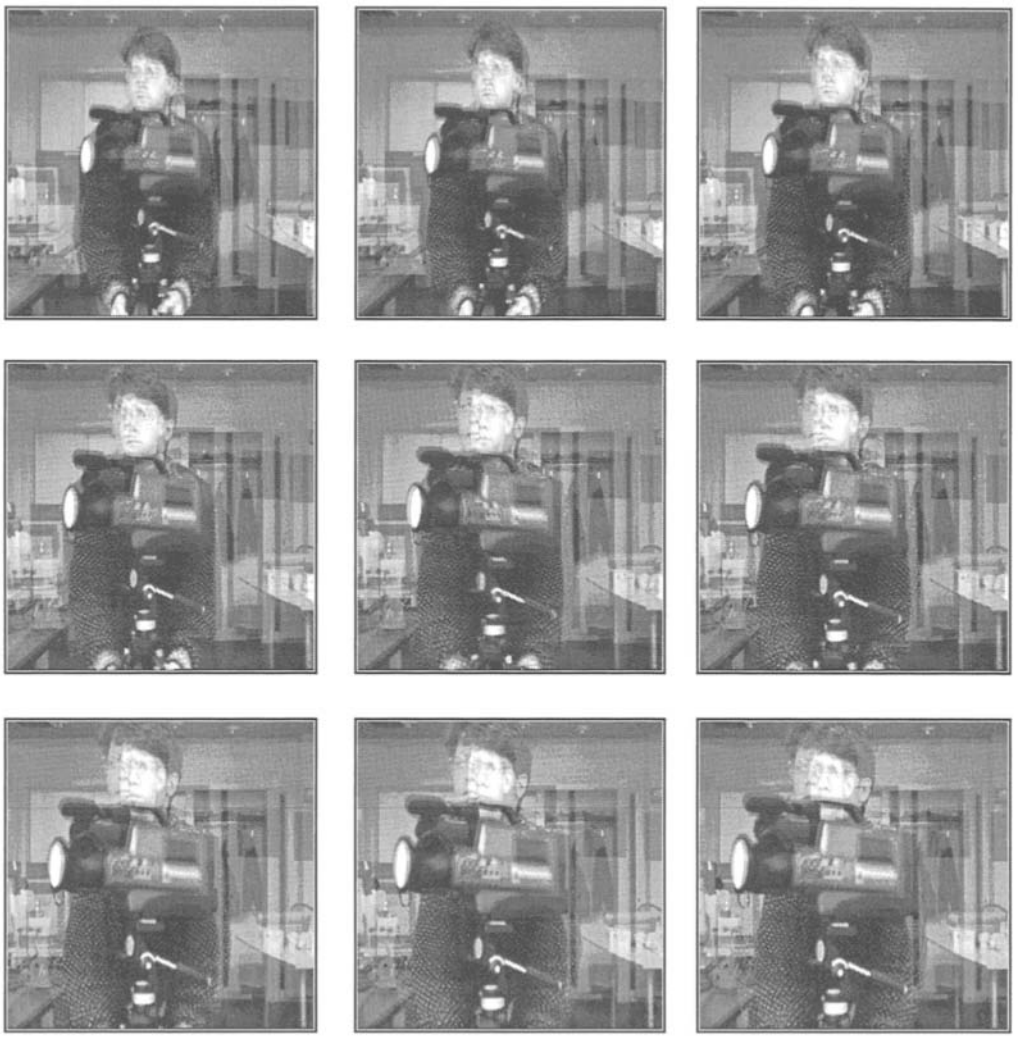

*Figure 18.* Top-left to Bottom-right. Binocular Tracking cooperating with vergence driven by a phase based disparity detection algorithm; images from left and right are superimposed and every tenth image is shown. The procedure is run in real time 25 frames/s and the fixation is maintained on the video camera in the superimposed image. The movement is mainly towards the head-eye system so as to depict the effect of disparity better.

To illustrate this, an experiment was performed using a phase-based disparity algorithm see e.g. (Sanger, 1988; Jepson and Jenkin, 1989; Calway et al., 1992). The basic concept here is to apply a convolution with a complex filter (e.g. a Gabor filter) to the left and right images. The linear proportionality of the phase difference and the shift in the images can then be used as a disparity estimate.

This algorithm can detect small disparities corresponding to depth changes that are comparable to the accelerations detectable by the tracking algorithm. This means that they can function together in parallel as to form a binocular tracking system, each process taking care of vergence and pursuit respectively. A tracking sequence resulting from this integration of the pursuing process with this vergence algorithm is illustrated in figure 18. The cameras are initially verged on the target and as the target starts moving towards the cameras they converge as to compensate for the emerging disparities.

It should be noted here that this algorithm needs no calibration. What is needed is that the vertical disparity

is known and that the relationship between pixel shift and angle change is approximately known, both of which are easily self calibrated. This fact is due to that the rotations of the cameras are performed around the projection centers and that the loop is closed with high frame rate. The details of the work on our implementation of a phase-based algorithm can be found in Maki et al. (1993). Here we have directly added the implementation to the set of our primary ocular processes.

## 10. Summary and Conclusions

Fixation is central to the human visual system; we are continuously fixating. Since we are looking at the world using monocular and binocular fixation, a machine vision system that interacts with the real world should also have both. Monocular and binocular fixations should be based on the same general schemes and representations. The fixation process itself cannot be static; it is needed precisely to handle the dynamism and varieties that are present in the visual world. Fixation is meaningful only if it is dynamic and we have shown that it is fully possible to implement systems capable of dynamic fixation in real time and in interaction with real objects.

These principles have been discussed in this article and we have shown how they can be implemented. Inspiration from the geometry and function of the human fixation mechanism has been our guide-line to applying the simple geometry of vergence. We have integrated monocular as well as binocular cues and processes with the goal of making a visual system that is capable of robust and reliable fixation.

The paper introduces a novel approach to modelling and implementing vergence in isolation from saccadic movements, in a dynamic scenario. Furthermore, the work covers the problems of sharpness and blur computations along with their application to active vision. In the same context a novel approach to compute depth from defocus was presented. Each single algorithm and different integration scenarios were accounted for in a series of experiments, both under static and dynamic conditions.

There are two future steps to take so as to accomplish a fully functional autonomously fixating agent, apart from enhancing existing algorithms and integrating new cues to the primary ocular architecture. The first step is to develop the work on the representational issues. This step requires a stable perceptual representation that in implementational terms simply means stabilizing the input images so that a *stable space* is perceived. The second step is to analyze the behavioral engine of fixation—the attentive mechanism, i.e. the strategy behind why the fixation is shifted and how. There is not much progress to expect on saccadic movements unless this issue is at least partially addressed.

## Acknowledgments

This work has been performed within a project sponsored by The Swedish Research Council for Engineering Science, TFR, with additional support from the basic research program in computer vision by The Swedish National Board for Industrial and Technical Development. This support is gratefully acknowledged.

We also thank Akihiro Horii who has implemented and studied depth from focus and blur, Atsuto Maki who has carried out the work on the phase based disparity detection and the members of the CVAP group for stimulating discussions.

## Notes

1. ... which is not claiming to be a complete model of human or primate oculo-motor system.
2. It is e.g. from the kinematic and image stability points of view interesting to rotate the lenses about the optical center. From other points of view, it might be also interesting to have other rotational schemes.
3. This is just a complementary example and not implemented.
4. Note that we try to be careful in using "binocular" and "stereo" here. In nature, having two eyes does not correspond to having stereo vision.
5. Note that the location is symmetric not the pattern.
6. Of course, this is true only if the two eyes are equally dominant. Otherwise, the angles follow the same proportionality as the dominance.
7. The classical experiments in physiology are cross references from Carpenter (1988) and Yarbus (1967).
8. This is not true in the case of primates. We have already mentioned that even pupil size is involved as a parameter in this context. The major task of the iris, however, seems to be regulation of the amount of light falling into the eyes.
9. In practice it is not so difficult to make our eyes do this kind of mistake.
10. See for example the many experiments on stereograms and fusion described in Julesz (1971).

## References

Abbott, A.L. and Ahuja, N. 1988. Surface reconstruction by dynamic integration of focus, camera vergence and stereo. *Proc. 2nd ICCV*, Tampa, FL., pp. 532–543

Alpern, M. and Ellen, P. 1956. A quantitative analysis of the horizontal movements of the eyes in the experiment of Johannes Müller. *American Journal of Ophthalmology*, 42:289–303

Alpern, M. 1957. The position of the eyes during prism vergence. *American Journal of Ophthalmology*, 57:345–353.

Calway, A.D., Knutsson, H., and Wilson, R. 1992. Multiresolution estimation of 2-D disparity using frequency domain approach. *Proc. BMVC*, pp. 237–246.

Carpenter, R.H.S. 1988. *Movements of the Eyes*, Pion press: London.

Clark, J.J. and Ferrier, N.J. 1988. Modal control of an attentive vision system. *Proc. 2nd ICCV*, Tampa, FL., pp. 514–523.

Coombs, D.J. 1992. Real-time gaze holding in binocular robot vision. Ph.D. Thesis, University of Rochester.

Fry, G.A. 1937. An experimental analysis of the accommodation-convergence relationship. *American Journal of Optometry*, 14:402–414.

Fry, G.A. 1939. Further experiments on the accommodation-convergence relationship. *American Journal of Optometry*, 16:325–336.

Horii, A. 1992. The focusing mechanism in the KTH head-eye system. TRITA-NA-P9215, Royal Institute of Technology, Computational Vision and Active Perception Laboratory, Stockholm.

Horii, A. 1992. Depth from defocusing. TRITA-NA-P16, Royal Institute of Technology, Computational Vision and Active Perception Laboratory, Stockholm.

Horn, B.K.P. 1989. *Robot Vision*, The MIT Press, p. 25.

Jarvis, R.A. 1976. Focus optimisation criteria for computer image processing. *Microscope*, 24(2):163–180.

Jepson, A.D. and Jenkin, M.R.M. 1989. The fast computation of disparity from phase differences. *Proc. CVPR*, pp. 398–403.

Julesz, B. 1971. Foundations of cyclopean perception. University of Chicago Press.

Knoll, H.A. 1949. Pupillary changes associated with accommodation and convergence. *American Journal of Optometry*, 26:346–357.

Krotkov, E.P. 1989. *Active Computer Vision by Cooperative Focus and Stereo*, Springer Verlag, 1989.

Luneburg, R.K. 1948. *Mathematical Analysis of Binocular Vision*, Princeton University Press: Princeton.

Maddox, E.E. 1886. Investigations in the relation between convergence and the accommodation of the eyes. *Journal of Anatomy*, 20:475–568.

Maki, A., Uhlin, T., and Eklundh, J.-O. 1993. Phase-based disparity estimating in binocular tracking. *Proc. 8th Scandinavian conf. on Image Analysis*, Tromsø, Norway.

Marg, E. and Morgan, M.W. 1949. The pupillary near reflex. The relation of pupillary diameter to accommodation and the various components of convergence. *American Journal of Optometry*, 26:183–198.

Marg, E. and Morgan, M.W. 1949. Further investigation of the pupillary near reflex; the effect of accommodation, fusional convergence and the proximity factor on pupillary diameter. *American Journal of Optometry*, 27:217–225.

Müller, J. 1826. *Zur Vergleichenden Physiologie des Gesichtssinnes*, C. Conbloch, Leipzig.

Pahlavan, K., Uhlin, T., and Eklundh, J.-O. 1992. Integrating Primary Ocular Processes. *Proc. 2nd ECCV*, Santa Margherita Ligure, Italy, pp. 526–541.

Pahlavan, K. and Eklundh, J.-O. 1992. A head-eye system-analysis and design. *CVGIP: Image Understanding, Special Issue on Active Vision*, Y. Aloimonos (ed.) 56(1):41–56.

Pahlavan, K., Uhlin, T., and Eklundh, J.-O. 1993. Active vision as a methodology. *Active Vision*, Y. Aloimonos (ed.), *Advances in Computer Science*, Lawrence Erlbaum, Hillsdale, NJ, pp. 19–46.

Pahlavan, K. 1993. Active robot vision and primary ocular processes. Ph.D. Thesis, TRITA-NA-P9316, Royal Institute of Technology, Computational Vision and Active Perception Laboratory, Stockholm.

Pahlavan, K. and Eklundh, J.-O. 1994. Mechatronics of active vision. *Mechatronics, Elsevier Science*, pp. 113–123.

Pentland, A. 1987. A new sense for depth of field. *Proc. IEEE Trans. PAMI*, pp. 523–531.

Pentland, A., Darrell, T., Turk, M., and Huang, W. 1989. A simple real-time range camera. *Proc. CVPR*, pp. 256–261.

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

Subbarao, M. 1988. Parallel depth recovery by changing camera parameters. *Proc. 2nd ICCV*, pp. 149–155.

Tenenbaum, J.M. 1970. *Accommodation in Computer Vision*, Ph.D. Thesis, Stanford University.

Yarbus, A. 1967. *Eye Movements and Vision*, Plenum Press: New York.

Zhang, W. and Bergholm, F. 1993. An extension of Marr's "Signature" based edge classification and other methods determining diffuseness and height of edges, and bar edge width. *Proc. 4th ICCV*, Berlin, pp. 183–191.