Vergence Micromovements and Depth Perception

Antônio Francisco *
CVAP, Royal Institute of Technology (KTH)
S-100 44 Stockholm, Sweden

Abstract

A new approach in stereo vision is proposed in which 3D depth information is recovered using *continuous vergence angle control* with simultaneous local correspondence response. This technique relates elements with the *same* relative position in the left and right images for a continuous sequence of vergence angles. The approach considers the extremely fine vergence movements (micromovements) about a given fixation point within the depth of field boundaries. It allows the recovery of 3D depth information given the knowledge of the geometry of the system and a sequence of pairs $[\alpha_i, \mathbf{C}_i]$, where α_i is the i^{th} vergence angle and $\mathbf{C}_i$ is the i^{th} matrix of correspondence responses. Due to its local operation characteristics, the resulting algorithms are implemented in a modular hardware scheme using transputers. Unlike currently used algorithms, there is no need to compute depth from disparity values; at the cost of the acquisition of a sequence of images during the micromovements. Experimental results from physiology and psychophysics suggest that the approach is biologically plausible. Therefore, the approach proposes a functional correlation between the vergence micromovements, depth perception, stereo acuity and stereo fusion.

The perception of the 3D-distance, depth, of objects using stereo images have been studied by many researchers for a long time. Some of these studies use vergence camera systems [1] integrating position control, image acquisition and depth processing on the modality of vision system named "active vision" [2]. Following this line of research, the present work analyses the correlation between the real time depth acquisition and the extremely fine vergence movements (micromovements) of the cameras about the fixation point. We assume that these movements are synchronized between the two cameras.

The *continuous vergence micromovements* differ from the vergence, translation and rotation movements used on the other methods to fixate the cameras on a new fixation point. The previous methods (using particular techniques as multi-resolution) compute some depth-map or depth directly from the acquisition of the left and right images at this fixation point, i.e., using two images and some stored information (estimation) about the depth-map they are able to infer the current depth at the correctly matched image points. Generally,

*Researcher at the National Institute of Space Research (INPE), São José dos Campos, São Paulo, Brazil. The support from the Swedish National Board for Industrial and Technical Development, NUTEK, is gratefully acknowledged. I would like to thank Prof. Ruzena Bajcsy and Prof. Jan-Olof Eklundh for the support to the development of this work as well as Kourosh Pahlavan, Akihiro Horii and Thomas Uhlin for valuable help when using the KTH head-eye system.

BMVC 1992 doi:10.5244/C.6.38

the strategies used is such methods have the search space for correspondence matches along epipolar lines. Therefore the depth is calculated using the disparity information between the left-right matched points and the geometry of the camera system.

The current approach uses neither epipolar lines nor disparities to calculate the depth of any 3D point. The depth is determined by the geometry of the camera system (mainly, the vergence angle) and by the relative position of a pixel with respect to the image plane. The procedure can be simply described as following: micromovements of two cameras occur about the fixation point. For each left-image point, on the left image plane, the vergence angle and the "correspondence response" of this point and the right-image point at the same relative position on the right image plane are stored. For each left-point, using these "correspondence response" signals and the camera geometry, the depth of the 3D points where the correspondence response reach the highest level are calculated.

Therefore, the approach is functionally different from the previous ones in the sense that the depth is calculated locally for each point (without searching epipolar lines) with the necessity of acquiring a sequence of images during the micromovements. The objective here is to clarify how to calculate depth using micromovements.

The paper covers the theoretical background, the experimental results and the biological support for depth acquisition from the vergence micromovents approach. The theoretical and simulations parts of this work were developed [3] during my stay at the General Robotics and Active Sensory Perception (GRASP) laboratory, University of Pennsylvania, USA. The experiments to validate the approach have been done using the KTH head-eye system [4].

1 The stereo vision system and the horopter

Each lens of the right and left camera is considered to be thin and *ideal*, in the sense that an object at a distance $\mathbf{d_{out}}$ (the *object distance*) from the principal plane has its image (with inverted direction) at distance $\mathbf{d_{in}}$ (the *image distance*) from this plane. The relationship of these two distances and the focal length of the lens is given by the *Gaussian Lens Equation:*

$$\frac{1}{f} = \frac{1}{d_{in}} + \frac{1}{d_{out}} \tag{1}$$

With respect to the camera platform, a symmetric fixation in the visual plane is assumed. Therefore the vergence angles of the two cameras have the same value α and the point being fixated is in the visual (horizontal) plane of the cameras. With this assumption, any camera torsion (about the axis connecting the lens center and the image plane center) is considered to be zero. According to the symmetric fixation model (figure 1) associated with each image plane there is a coordinate system with its origin at the image plane center. The lens centers are separated by a baseline $\mathbf{b}$. These coordinate systems define the left projection (x_{pl}, y_{pl}) and the right projection (x_{pr}, y_{pr}) of a point in space. To identify the 3D position $(X_o, Y_o, Z_o)^T$ of a point $\vec{P}_o$ in space a global coordinate system xyz is used (figure 1).

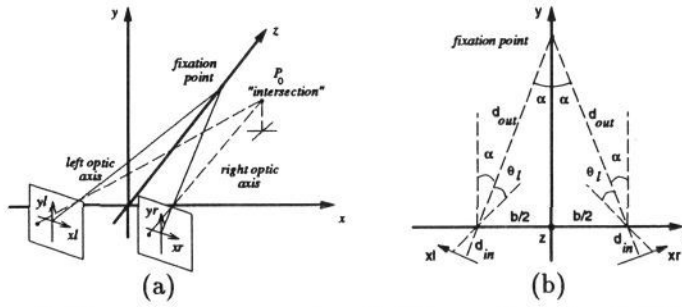


Figure 1: Stereo camera geometry: (a) Perspective view (b) Top view

In order to adopt the same terminology used in the human vision field we will review some concepts from the physiology and psychophysics sciences. The **horopter** defines the set of points in space for which the binocular disparity is zero [5]. The **point horopter** is the locus of zero disparities for the point stimulus where *both* horizontal and vertical disparities are zero. (There has been a considerable amount of confusion in the literature caused by the laxity in defining the horopter [5].)

We consider the ideal point horopter composed of points with zero horizontal and vertical disparities for the symmetric fixation in the visual plane, with any position, torsion and optical aberrations assumed to be absent. As described in [5], any point off the *point horopter* in space (off-axis points) projects to the two image planes with horizontal and vertical disparities. With vergence, the points at the distance corresponding to the point horopter would nullify the horizontal disparity. Note that in the ideal case nothing can be done to nullify the vertical disparity produced by off-axis points being necessarily closer to one eye than the other, with a resulting difference in the projection angle in the two eyes. The present analysis concerning the horopter is based on *zero horizontal* disparity.

For the *zero horizontal disparity* case we can define a 3D "intersection" point of the left and right optic axes (same x and z and different ys) passing through the same corresponding element $(x_p, y_p)^T$ on the image planes coordinate systems. The coordinate $(x_\iota, y_\iota, z_\iota)^T$ of this "intersection" point [3] is:

$$\left(\frac{b}{2}\frac{\tan(\alpha+\theta_l)-\tan(\alpha-\theta_l)}{\tan(\alpha+\theta_l)+\tan(\alpha-\theta_l)},\quad -\frac{y_p}{d_{in}}\frac{b}{2}\frac{cos^2(\theta_l)}{\sin(\alpha)},\quad \frac{b}{\tan(\alpha+\theta_l)+\tan(\alpha-\theta_l)}\right)^T \tag{2}$$

where,

$$\alpha = \arctan\left(\frac{b}{2d_{out}}\right),\quad \theta_l = -\arctan\left(\frac{x_p}{d_{in}}\right),\quad \text{and (eq. 1)}\quad d_{in} = \frac{d_{out}f}{d_{out}-f} \tag{3}$$

The above equation for y_ι was deduced considering the average of the left and right y coordinates of the left and right optic axes at the "intersection" point. This assumption includes an error in the present analysis. In order to evaluate the dimension of this error with respect to the length of the photoreceptor element, the difference between the projections on both image planes of a point $\vec{P_o}$ in $(x_\iota, y_\iota, z_\iota)^T$ (eq. 2) is analyzed. Note that the projections are determined by the intersection of the right and left optic axes, passing through the left and right optic lens center respectively, with the image planes. Let us denote the intersection of the left optic axis with the left image plane as $(x_{pl}, y_{pl})^T$ and the intersection of the right optic axis with the respective

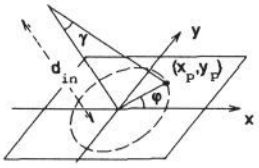


Figure 2: Polar coordinate system of the image plane

image plane as $(x_{pr}, y_{pr})^T$. It is possible [3] to define the following euclidian *projections deviation* (*dev*) for a given point $\vec{P_o}$ as:

$$dev = \sqrt{(x_{pl} - x_{pr})^2 + (y_{pl} - y_{pr})^2} \tag{4}$$

where,

$$x_{pl} = -d_{in} \tan(+\arctan(\tfrac{\frac{b}{2}+X_o}{Z_o}) - \alpha), \quad y_{pl} = \tfrac{-Y_o d_{in} \sin(\alpha+\theta_l)}{\cos(\theta_l)(X_o+\frac{b}{2})} \tag{5}$$

$$x_{pr} = -d_{in} \tan(-\arctan(\tfrac{\frac{b}{2}-X_o}{Z_o}) + \alpha), \quad y_{pr} = \tfrac{Y_o d_{in} \sin(\alpha+\theta_r)}{\cos(\theta_r)(X_o-\frac{b}{2})} \tag{6}$$

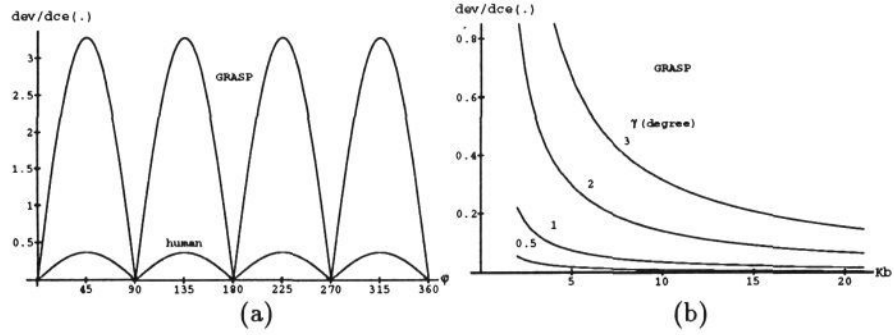
$$\theta_l = -\arctan(\tfrac{x_{pl}}{d_{in}}), \qquad \theta_r = \arctan(\tfrac{x_{pr}}{d_{in}}) \tag{7}$$

The *dev* analysis is done considering the object distance as a multiple of the baseline ($d_{obj} = k_b b$) and the image planes mapped by a polar coordinate system (figure 2). Having deduced all needed equations for the *dev* analysis, it is time to show some simulated results about the human visual system and the GRASP platform system. The procedure to accomplish the *dev* analysis can be synthesized in the following simulation steps:

- for a given: k_b, γ, φ, f , b ; compute: d_{obj} , α , d_{in} (eq. 3), x_p, y_p (figure 2), $(x_\iota, y_\iota, z_\iota)^T$ (eq. 2),
- using $\vec{P_o} = (x_\iota, y_\iota, z_\iota)^T$, compute: x_{pl}, y_{pl}, x_{pr}, y_{pr} (eq.s 5 to 7) and then *dev* (eq. 4).
- plot: the results of the *dev* normalized with respect to the distance between centers of adjacent photo-receptor elements *dce*. Note that *dce* is a constant for most machine vision systems (*dce*(.)) and is a function of γ for the human visual system (*dce*(γ)) [3].

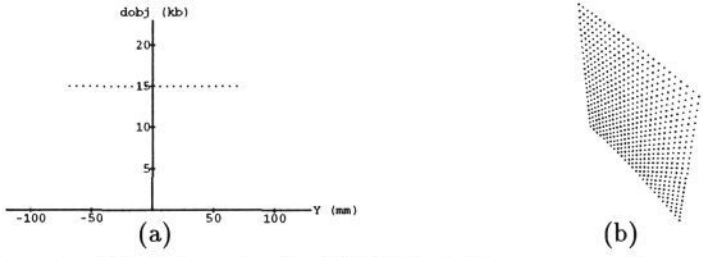
The results of the simulation shown in Figure 3.a imply that the highest *dev* occurs when $\varphi = k$ 90 degree + 45 degree (k = 0,1,..). Therefore all other simulations are done with the value of φ equal 45 degree. A conclusion from Figure 3.a, is that the normalized *dev* is smaller for the human visual system since *dce*(γ) increases on the periphery. This characteristic implicit in the human visual system tends to diminish *dev*. Another feature of the human visual system that tends to diminish *dev* is the known difference between nasal and temporal retina eccentricity (if nasal is larger than temporal) for every pair of corresponding points. This difference in eccentricity could explain the deviation of the empirical horopter [5] from the V-M circle as well as the necessity to diminish *dev*.

For the GRASP system we want to know how *dev* varies with object distance. From the analysis of Figure 3.b we see that *dev* decreases with object



(a) (b)

Figure 3: Normalized deviation as a function of: (a) φ ($\gamma = 2^o$, b = 65 mm, f = 17 mm, $d_{obj} = 2b$), (b) k_b ($d_{obj} = k_b b$, $\varphi = 45^o$, b = 128 mm, f = 65 mm)



(a) (b)

Figure 4: Point horopter for GRASP platform system (b = 128 mm, f = 65 mm, $d_{obj} = 15$ b, dce(.) = 30.0 10^{-3} mm): (a) Top view, (b) Perspective view

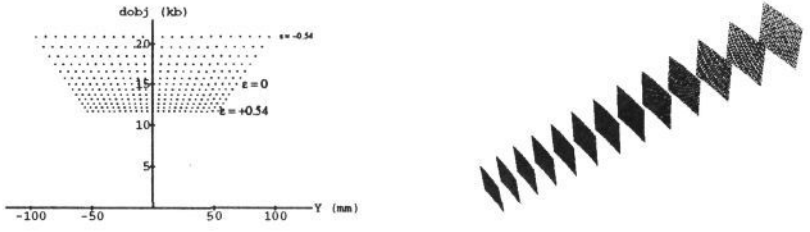
distance, therefore the next set of simulations are done with $d_{obj} = 15b$ which ensures a small *dev* for $\varphi = 45$ degree. The *dev* has been investigated for its maximum value, in spite of the zero value of this deviation on the image planes coordinate axes ($\varphi = k$ 90 degree, k = 0,1,..) for any value of γ, d_{obj}, b and f.

Having shown that the normalized *dev* is very small when d_{obj} is greater than fifteen times the length of the baseline, the *point horopter* is plotted using equation 2 for this value of d_{obj}. Almost all the parameters of the above equations can be computed directly (like d_{in} and α) from the defined values of d_{obj}, b and f. The only two parameters that do not have a defined range are x_p and y_p. In the present paper, 80 photo-receptor elements are used as the distance from the image plane center to the periphery of the workspace being analyzed. Therefore, a *square workspace* of side size equal to 160 pixels centered in the image plane is assumed. This range of x_p and y_p gives the point horopter plotted in Figure 4 for the GRASP platform system. It can be seen that the point horopter is a surface in space.

2 Micromovements

The shape of the point horopter has been analyzed for a given vergence angle calculated from the object distance under fixation. The main analysis now is conducted for a number of vergence angles α_i about the *fixation point in the visual plane*, described by the following equation:

$$\alpha_i = \arctan(\frac{b}{2d_{obj}}) + \varepsilon_i \tag{8}$$



(a) (b)

Figure 5: Set of point horopter surfaces for GRASP platform system ($\varepsilon_i \in [-0.54^o, 0.54^o]$): (a) Top view, (b) Expanded perspective view

where ε_i is a small angle increment (positive or negative) that describes the micromovement about the *fixation point* (first part of equation 8). The set of α_i about a given fixation point determines a complete *micromovement cycle.*

Figure 5 shows the locus of "intersection" points in 3D space of the GRASP platform system for a given micromovement cycle. The surfaces shown correspond to the set of point horopter surface generated for each vergence angle α_i. As can be seen in Figure 5, the locus of all the "intersection" points form a volume in the 3D space. Therefore, any object inside this volume can have its depth measurements determined by the response of a *local correspondence operator* to the *continuous vergence angle control.* Remember that this operator relates elements (image plane points) with the *same* horizontal and vertical distance from the center of the left and right image planes. It is possible to use a local correspondence operator since we assume that d_{obj} is greater than fifteen baselines, implying a small *dev* (see previous section).

The errors in stereo (along z axis) with the present approach are due the vergence angle quantization (angle steps fixated by ε_i). These errors differ from the quantization errors due to discrete photo-elements in cameras, that are a common characteristic of other stereoscopic methods. As described in [6], the errors due the photo-receptor quantization are significant and increase with the distance from the object to the cameras system. The present approach allows us to overcome the photo-receptor quantization limitation by using a sequence of pairs $[\alpha_i, \mathbf{C}_i]$, where α_i is the i^{th} vergence angle and $\mathbf{C}_i$ is the i^{th} matrix of correspondence responses. Although the present analysis considers only the micromovements in the visual plane (horizontal micromovements), the human eye system performs micromovements in a vertical plane including the visual axis as well as rotations about the visual axis itself [7].

3 Biological support of the micromovements

The following discussion of eye-movements according to physiological and psychophysical experiments is offered as a working hypothesis, useful for the understanding the role of the micromovements on depth perception. Physiological results [8, 7] show that the human eye performs fine movements during the process of fixation on a single point, which are collectively called *physiological nystagmus.* Physiological nystagmus is composed of three different kinds of movements: (1) high-frequency tremor, (2) slow drifts, and (3) rapid *binocular flicks.* The drift and flick movements occur in opposing directions and produce convergence/divergence waves of the eyes on a similar way as the micromove-

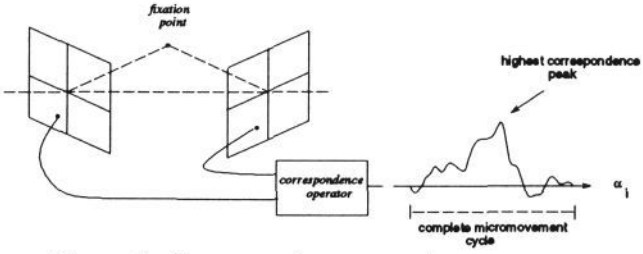


Figure 6: Correspondence operator response

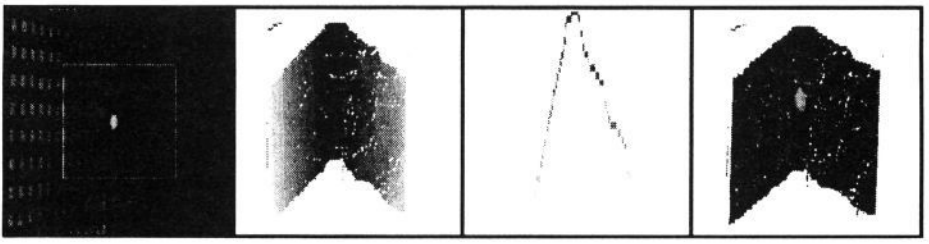
ments studied in previous sections.

Assuming the vergence micromovements mechanism as the basis of the depth perception, it is easy to understand the phenomena of *stereoacuity* (depth or stereoscopic acuity, stereopsis). As well described in [8, 5], it is almost incredible that most observers under normal conditions can discriminate a difference in depth corresponding to an *angular disparity* (interocular disparity) of about 10 arc sec. The best values reported in the literature have been obtained by the apparatus called the *Howard-Dolman apparatus*, devised by Howard in 1919. The best observers achieve a 75% discrimination level close to 2 arc-seconds in that experiment. The most incredible fact is that this disparity value is much smaller than the distance between the cones' centers at the *central part* of the *fovea* ($\approx$ 22 arc sec).

We suggest that the high sensitivity to slight disparity can be explained by the correlation between depth perception and the vergence eyes micromovements and not by the capacity of the human visual system to spatially detect disparity on the retinas. Therefore the idea of an *angular disparity* that can be detected *spatially* by the visual system is substituted by a *local approach* where the human visual system determines the depth values by the highest peak of correspondence response (figure 6) during a complete *micromovement cycle* (section 2). The highest peak of correspondence occurs when there is no spatial disparity between the left and right stimulus of elements with the same relative position on both retinas, i.e., when the spatial disparity is cancelled for a given vergence angle.

Another phenomenon that can be explained by the present approach is known in the literature [8] as *Panum's fusional area*: the range of interocular disparities within which objects viewed with both eyes on corresponding retinal regions appear single. This area is such that *fusion* occurs, only one dot is seen, when two points that are perceived in different eyes fall closer together in the combined view. Note that these two points can be seen through an uncrossed (left and right optic axis do not cross) or crossed disparity. The classical static limits for Panum's area, the mean crossed to uncrossed range of horizontal disparities, is reported as being 14 arc min. The experiments described in [9] support the existence of *binocular fusion* as a *unique* category of sensory performance, disconfirming several non fusional explanations of single vision. While the range of binocular disparities allowing fusion (Panum's fusional horizontal diameter) is typically in the region of 14 arc min, stereoscopic depth can be perceived from a disparity 500 times smaller.

In the present approach, the phenomena of binocular fusion and stereoscopic depth are assumed to be supported by the mechanism of vergence eye micromovements about a fixation point. In this way, the fusion area dimension



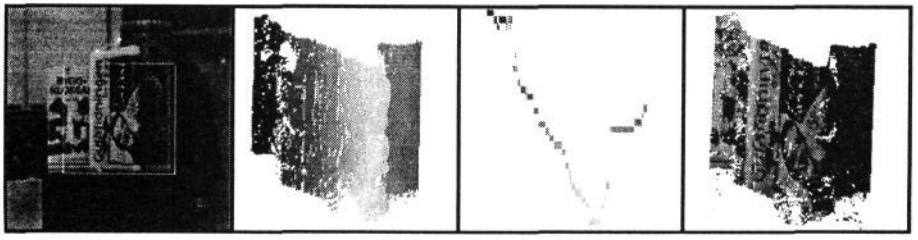
(a) (b) (c) (d)

Figure 7: Two planes experiment (33 vergence steps): (a) original image, (b) perspective view of the acquired object, (c) horizontal cut of the acquired object, (d) perspective view pasted with real grey value from original image

is determined by the range of a complete *micromovement cycle* (section 2). It is important to point out that the *classical* value of the Panum's fusional *horizontal radius* (average of the crossed and uncrossed disparities), 7.0 arc min, coincides with the micromovement range value described in [7]. Note that the Panum's fusional horizontal radius must be compared to the total range of a *monocular* micromovement reported in [7] to be coherent to both definitions. In the present analysis the *vertical* fusion radius is not considered since this radius follows the monocular spatial resolution limit of the retina [10]. As a conclusion, the "real" binocular fusion is assumed to occur only between cells adjacent on the horizontal axis of the retinas, and that binocular vertical fusion is a result of the monocular fusion mechanism.

4 Experimental results

In order to validate experimentally the micromovements approach, a practical implementation was done using the KTH head-eye system. This active vision system is composed of several motors, two cameras and two camera lens controllers. The system is connected on the VME-bus of a SUN-SPARC station via a transputer board. Our main control was over the vergence motors, image acquisition, zoom and focus. At the beginning, we do not use the transputer board to execute the algorithms. Instead of that we did the experiments acquiring and storaging a sequence of images pairs in the SPARC station for further processing. The two experiments that will be described below consist of an object located in front of the head-eye system around 3000 mm from the baseline. The set up is done by choosing a value for the focal length (zoom) and adjusting manually the focus and the initial fixation point over the central part of the object surface. These values of zoom and focus are kept constant during the experiment. A program makes the sweeping of the object by changing the vergence angle between the initial vergence value and a final value determined by the number of steps specified. For each vergence angle, two images (from the left and right cameras) are stored on the SPARC station. After the acquisition of the desired number of images pairs, we compute the correspondence response for every pixel (x,y) on the acquired left and right images, using the



(a) (b) (c) (d)

Figure 8: The box-plane-cylinder experiment (41 vergence steps): (a) original image, (b) perspective view of the acquired object, (c) horizontal cut of the object, (d) perspective view pasted with real grey value from original image

following correlation operator:

$$Corr(x,y) = \frac{E[Left(x,y)Right(x,y)] - E[Left(x,y)]E[Right(x,y)]}{\sigma[Left(x,y)]\sigma[Right(x,y)]} \quad (9)$$

where,

$$E[X(x,y)] = \frac{\sum_{i=-w/2}^{w/2}\sum_{j=-w/2}^{w/2} X(x+i,y+j)}{2(w+1)}, \quad (\sigma[X])^2 = E[X^2] - (E[X])^2$$

For every pixel, using the vergence angle α_i that gives the highest correspondence response for that pixel and using (8) we compute the object depth d_{obj} for that pixel.

The following parameters are common in both experiments: b = 200 mm, f = 30 mm, d_{obj} = 15 b, the work window size of 200x200 pixels, the operator size of 21x21 pixels and the vergence step resolution of 26 arc sec (imposed by the head-eye system). Our first experiment was done using two vertical planes as the object being viewed. In Figure 7.a the left image of the object before the vergence sweeping is shown. Figure 7.b shows the perspective view of the acquired object after the vergence sweeping and the use of equation 8. The darker grey patches of the object are farther from the baseline than the lighter grey patches. On Figure 7.c, a horizontal cut of the acquired object is shown permitting us to have the correct idea of the object being viewed. The last picture is the perspective view pasted with the real grey values instead of the depth represented grey values. The second experiment is shown in Figure 8. The object is composed of a box (left side of the object) a cylinder (right side) and an inclined newspaper. Figure 8.c gives the notion of the object used.

The step-shape wave seen on Figures 7.c and 8.c is a consequence of the vergence step resolution. The processing time using the previous scheme was about one hour for the processing of the entire 200x200 pixel being viewed. Actually the correspondence operation is being executed on the transputer board. The entire image was split in four transputers before the correspondence operation. Using this new scheme the experiment took around 20 seconds. We are not using the entire power of our transputer board since we did not have time enough to implement it. In spite of the great improvement using the transputer board our goal was not reach yet, since we want to process depth at the frame rate.

5 Conclusion

The *continuous vergence micromovements* approach permits to overcome the physical limitation of the photo-receptor dimension (CCD element or cone) on the depth perception. Moreover, there is no need to compute depth from disparity values since the disparity is cancelled by the vergence micromovements. Note that the *stereoscopic matching problem* still exists, since there is the possibility to have two or more correspondence peaks with similar values for an element of the correspondence matrix.

The highlight of this new approach is the vergence micromovements as a mechanism to nullify the disparity between the left and right visual stimulus at the same retina locus. Therefore, the concept of a "neural structure spread spatially" in the visual system to perceive depth via measurement of disparity is substituted by a "neural structure connected locally with the neighborhood" of each retina locus.

References

[1] E. P. Krotkov. *Exploratory visual sensing for determining spatial layout with an agile stereo camera system.* PhD thesis, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA, 1987.

[2] R. Bajcsy. Active perception vs. passive perception. In *Proc. Workshop on Computer Vision*, pages 55–59, Bellaire,MI, October 1985.

[3] A. Francisco. The role of vergence micromovements on depth perception. Technical Report MS-CIS-91-37, GRASP LAB, CIS, University of Pennsylvania, Philadelphia, PA, USA, 1991.

[4] K. Pahlavan and J.O. Eklundh. A head-eye system - analysis and design. In *Computer Vision, Graphics, and Image Processing: Image Understanding*, page (To appear.), July 1992.

[5] C. M. Schor and K. J. Giuffreda. *Vergence eye movements: basic and clinical aspects.* Butterworth, 1983.

[6] F. Solina. Errors in stereo due to quantization. Technical Report MS-CIS-85-34, GRASP LAB, CIS, University of Pennsylvania, Philadelphia, PA, USA, 1985.

[7] R. W. Ditchburn. Eye-movements in relation to retinal action. *Optica Acta*, 1(4):171–176, 1955.

[8] J. W. Kling and L. A. Riggs. *Experimental psychology.* Holt, Rinehart and Winston, Inc., 1971.

[9] T. Heckmann and C. M. Schor. Panum's fusional area estimated with a criterion-free technique. *Perception & Psychophysics*, 45(4):297–306, 1989.

[10] C. Schor, I. Wood, and J. Ogawa. Binocular sensory fusion is limited by spatial resolution. *Vision Res.*, 24(7):661–665, 1984.