Application of a Novel Neural Approach to 3D Gaze Tracking: Vergence Eye-Movements in Autostereograms

Kai Essig[1], Marc Pomplun[2] and Helge Ritter[1]

[1]Neuroinformatics Group, Faculty of Technology, Bielefeld University,
P.O.-Box 10 01 31, 33501 Bielefeld, Germany
Email: (kessig, helge)@techfak.uni-bielefeld.de

[2]Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Boulevard, Boston, MA 02125-3393, USA
Email: marc@cs.umb.edu

Abstract

Vergence eye-movements occur not only in real environments, but also in virtual 3D environments. Autostereograms can cover large visual angles without requiring vergence beyond natural parameters and are thus well-suited for the investigation of vergence movements in virtual 3D images. We developed an anaglyph-based 3D calibration procedure and used a *parametrized self-organizing map* (PSOM) to approximate the 3D gaze point from a subject's binocular eye-movement data. Besides analyzing the general pattern of vergence movements in autostereogram images, the present study examined the influence of image granularity on these movements. Unlike previous research on random-dot stereograms, we found substantially overshooting convergence eye-movements, especially for medium granularities. Moreover, divergence movements were completed more quickly for coarse than for fine granularities. Results are discussed in the context of granularity effects on autostereogram perception and the dissociation between convergence and divergence eye-movements.

Vergence Eye-Movements in Autostereograms

In everyday life, we employ vergence eye-movements to successively fixate objects at different distances. It is interesting to note that these movements are also produced in virtual 3D environments. Some work has been done on the analysis of vergence movements occurring while viewing *Random-Dot Stereograms* (*RDS*). For *RDS* two slightly different images for the left and the right eye seen through a special device, a so-called stereoscope, lead to the perception of 3D information. Mowforth, Mayhew and Frisby (1981) found that *RDS* are perceptually filtered by spatial frequencies. Low frequencies enable the 3D perception of *RDS* with larger disparities and lead to faster vergence movements than do high frequencies. Furthermore, vergence velocities were found to be faster for convergent than for divergent eye movements.

In another study, Rashbass and Westheimer (1961) found reaction times of about 160 ms for both convergent and divergent eye movements in response to suddenly introduced target vergence changes. The authors reported start velocity varying with stimulus amplitude, and vergence reaching asymptotically its final level after approximately 800 msec without any overshoots or a prolonged period of oscillations. They also found turn-abouts in the response before the target vergence change

reaches zero, showing the anticipatory behavior of the eye-vergence system.

For our experiment, however, instead of *RDS* we used *Single Image Random Dot Stereograms (SIRDS)*, also called *autostereograms*. The difference between *RDS* and *SIRDS* is that in *SIRDS* the two slightly different images of the *RDS* are combined into a single image. No stereoscope is necessary to perceive the 3D information in a *SIRDS*. To achieve the 3D perception of a *SIRDS*, observers have to combine the information of the two images for the left and the right eye to one depth perception by focusing a point before (*cross-eyed method*) or behind (*wide-eyed method*) the image plane, whereby our preliminary studies (Essig, 1998) demonstrated that the later method leads to a more stable perception of the depth information and was therefore used for the reported experiments. Once the observers perceive the 3D scene, they may move their gaze to any position in the image without losing the 3D impression. Most important in the present context, *SIRDS* can cover large visual angles without requiring vergence beyond natural parameters, which makes them appropriate stimuli for studying vergence movements in virtual 3D images.

In our natural environment, and when looking at *RDS*, we move our eyes inward (*convergence*) to inspect near objects and outward (*divergence*) to inspect distant objects. An obvious question is: Do these vergence movements also occur during the observation of autostereogram images? Belopolskii and Logvinenko (1994) found that vergence movements are not *necessary* to get the 3D perception of an autostereogram. However, in recent experiments (Essig, 1998) we found that in fact vergence eye-movements are executed during the examination of autostereogram images. These results encouraged us to conduct further experiments investigating stereogram parameters that are likely to influence vergence movements.

A distinguishing feature of autostereogram images is their grain size (granularity). Figure 1 shows examples for autostereogram images with small (upper panel) and big grain sizes (lower panel). The reported experiment used autostereogram images with different grain sizes to systematically investigate the influence of this parameter on the vergence movements. According to Mowforth et al. (1981), overshoots and oscillations should be more likely for autostereogram images with coarser granularities, i.e. lower spatial frequencies, which, however, is in-
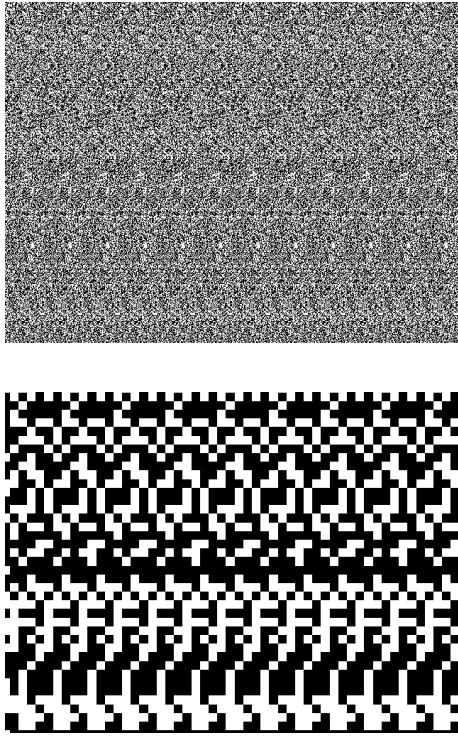
Figure 1: Autostereogram images with granularities 1 (upper panel) and 12 (lower panel).

consistent with the findings of Rashbass and Westheimer (1961). For a clarifying analysis of this issue, we measured vergence eye-movements with a modern binocular eye tracker with a sampling rate of 250 Hertz. For increased precision of vergence and 3D gaze-position measurement, we developed and applied a novel neural-net based calibration interface, which is described in the following section.

A Neural Approach to 3D Gaze-Point Calculation

During the examination of autostereograms shown on the computer screen, the intersection of the viewing axes is in general in front of or behind the screen plane, depending on the 3D gaze position. As a consequence, the correct coordinates for the actual gaze position are different from the screen coordinates which the eye tracker provides, because the system uses a 2D calibration to calculate the relation between the pupil positions and the gaze point on the screen. We tackled this problem by developing a 3D calibration that precedes the experiment and uses a *parametrized self-organizing map* (PSOM) to approximate the 3D gaze point from a subject's binocular eye-movement data.

Since the experimental conditions change from subject to subject (e.g. subjects have physical differences, their gaze characteristics are individually different, and the eye-tracker setup varies across sessions), it is clearly advantageous to use a method which is able to "learn"

these individual parameters. Additionally, we have to take into consideration that the mapping from pupil to screen coordinates is non-linear. Hence neural nets are suitable for the solution of these problems, because they are able to "learn" nonlinear functions. Kohonen (1990) developed the so-called *self-organizing maps* (SOMs), which could learn the correct mapping from the pupil to the screen coordinates. However, SOMs have two disadvantages: They supply only the position of the most stimulated neuron in a "neuron lattice" instead of a continuous output, and they usually require thousands of training examples. Therefore, we use a variant of SOMs, namely a so-called *parametrized self-organizing map* (PSOM) (Ritter, 1993). This variant does not have the disadvantages mentioned above, because it provides the demanded continuous output and gets only some selected input/output pairs as parameters, i.e. the coordinates of the calibration points and the related positions of the pupils measured by the eye tracker.

A 2D version of the PSOM was already used in (Pomplun, Velichkovsky & Ritter, 1994), reducing the average calibration error to about 30% of its previous value. In opposition, our "new" PSOM does not only enhance the accuracy of measurement, it also approximates the subject's 3D gaze position from the two 2D coordinates of the eye tracker system.

A PSOM can be considered as a recursive neural net that realizes a distinct, mostly multi-dimensional projection $\mathbf{f}$. The input data of this projection consist of the "correct" coordinates of 27 calibration points $k \in \mathbf{A}$ (standardized in the interval from 0 to 2 by the PSOM), where $\mathbf{A} = \{k_{xyz} | k_{xyz} = x\hat{e_x} + y\hat{e_y} + z\hat{e_z}; x, y, z = 0...2\}$
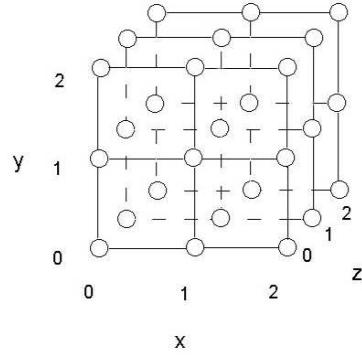

Figure 2: The calibration points in the virtual 3 x 3 x 3 cuboid. For the PSOM the coordinates are standardized in such a way that they only have the values 0, 1, and 2.

(see Figure 2) arranged in a 3 x 3 x 3 grid and presented to the subjects during the calibration procedure. In order to create a virtual 3D perception, the calibration grid was presented as an anaglyph image, viewed through red-green glasses. The calibration points were drawn on the planes of a cuboid, where the planes before and behind the screen plane formed its surfaces. This made it perceptually easier for the subjects to locate the stimuli on one of the three planes of the cuboid. For each of these

27 calibration markers, we have the associated gaze coordinates measured by the eye tracker and the x-divergence between the gaze positions of the left and right eye on the screen calculated from this data. Thus, the reference vector is $w_k = (x_{plk}, y_{plk}, x_{prk}, y_{prk}, x_{dk}) \in \mathbf{R}^5$. x_{plk} is the x-coordinate for the left eye, and y_{plk} the corresponding y-coordinate for the left eye belonging to the vector w_k. For the right eye the corresponding values are x_{prk} and y_{prk}. The fifth element of w_k results from $x_{dk} := x_{prk} - x_{plk}$.

We introduce the divergence (x_{dk}) as the fifth dimension of w_k because the z-coordinate of the 3D gaze-position mainly depends on this divergence. Since the differences in the divergence are smaller than those in the x- and y-directions, the divergence has to be weighted by a specific factor. This method leads to a faster termination of the PSOM-calculations. Hence, every fixation point $k \in \mathbf{A}$ is associated with a corresponding reference vector $w_k \in \mathbf{R}^5$. With the reference vectors w_k we could already construct a SOM which could enable the projection from the 2D coordinates of the system to the "correct" 3D coordinates. This SOM would only enable a crude approximation to the real 3D coordinates, because it could only choose one of the calibration points as a possible output. In the present context, however, we are looking for a projection which can interpolate between the vectors w_k. This projection can be done by a PSOM.

The desired interpolation function $\mathbf{f}$(s) can be constructed from the superposition of a suitable number of simpler basis functions H(.,.) as follows:

$$\mathbf{f}(s) = \sum_{k \in A} H(s, k) w_k \qquad (1)$$

The values of the basis functions are within the interval [0,1] depending on different values s, so that the coordinates of a calibration point, which is close to the desired gaze position, is weighted strongly (near 1) and points which are far away are weighted weakly (near 0). The basis functions $H : \mathbf{R}^3 \times A \to \mathbf{R}$ have to comply with the requirement

$$H(s, k) = \delta_{s,k} \qquad \forall \ s, k \in A \qquad (2)$$

where δ represents the Kronecker symbol. It is defined as: $\delta_{ij} = \begin{cases} 1 & : & i = j \\ 0 & : & i \neq j \end{cases} \quad \forall \ i, j = 0, 1, 2$
(in this special case). This ensures that

$$\mathbf{f}(s) = w_s \qquad \forall \ s \in A$$

Thus, it is guaranteed that the interpolation function passes through the given points. But how can we choose suitable functions H that obey equation (2), that are smooth and simple to handle?

One convenient solution is to make a product ansatz and to derive the suitable function by combining three 1D functions (one in each case for every coordinate direction x, y, and z). The new 1D functions must then have the property $H^{(1)} : \mathbf{R} \times \{0, 1, 2\} \to \mathbf{R}$:

$$H^{(1)}(q, n) = \delta_{q,n} \quad \forall \ q \in \mathbf{R}, n \in \{0, 1, 2\} \qquad (3)$$

Because n has only three possible values (0, 1, and 2), it is sufficient to choose an account of three basis functions $\mathbf{R} \to \mathbf{R}$. For this purpose, polynomials of second degree are especially suitable, because they have no redundant degrees of freedom.

Now we have built a function which projects 3D coordinates onto 2D gaze-positions for both eyes. Our final aim is to find a function which does just the opposite though, namely to approximate the 3D gaze-position of the subject from the system's 2D measurements. Therefore, we have to calculate the inverse function $\mathbf{f}^{-1}$ of $\mathbf{f}$. The non-linearity of $\mathbf{f}$ forces us to use a numerical procedure. Thus we have to create an error function $E(s)$, which is defined as:

$$E(s) = \frac{1}{2}(\mathbf{f}(s) - \mathbf{f}_{et})^2 \qquad (4)$$

i.e. the deviation of the pupil coordinates provided by the eye tracker ($\mathbf{f_{et}}$) from the eye tracker data calculated by the PSOM $\mathbf{f(s)}$ for the actually assigned two screen points.

If this difference exceeds a specified threshold, the coordinates of these points are modified in an iterative gradient-descent procedure until the results of Equation (1) closely approximate the actual eye tracker data provided by the system:

$$s(t + 1) = s(t) - \epsilon \cdot \frac{\delta E(s)}{\delta s} \qquad , \text{with} \ \epsilon > 0 \qquad (5)$$

This means that the iteration process stops if $E(s(t))$ falls below the threshold, which we should adapt to the screen resolution. In this way, an exact 3D gaze position of the subject during the examination of a 3D stimulus is assigned to the 2D eye tracker data for the left and right eye.

We conducted an experiment in which the accuracy of the PSOM gaze-point calculation is compared to one provided by a geometrical solution. Subjects had to visually track a dot that appeared at positions of a 4 x 4 x 4 grid in 3D space in a random sequence. In the geometrical method the equations for the right and left visual axes are calculated on the basis of both the measured gaze-positions on the screen and the subjects' (assumed) constant head position. The gaze point is determined as the point where the visual axes are closest to one another. It results as the center of the shortest straight link between the visual axes.

The results show that the neural net calculates gaze positions from the eye tracker data which are nearly 46% closer to the actual gaze positions than those of the geometrical solution. The average total error for all subjects, separated for the individual coordinates, is shown in Table 1 (the values for the geometrical method and the neural net show the average total error and the standard error).

It is obvious that the z-error is always higher than the $x-$ and $y-$errors, because the z-coordinate is much more sensitive to small changes in the binocular gaze position than the $x-$ and $y-$coordinates. We also found that the measurement errors decrease from the back to the front

Table 1: Average total error for individual coordinates.

coordinate	both methods average total error	geometrical method average total error	neural net average total error
x	0.97cm	1.41cm ± 0.04cm	0.52cm ± 0.05cm
y	1.03cm	1.24cm ± 0.05cm	0.82cm ± 0.05cm
z	4.16cm	5.79cm ± 0.36cm	2.53cm ± 0.11cm

plane. The neural net compensates the errors in the back plane better than the geometrical method, because small changes in the gaze position at the back plane lead to severe inaccuracies. Obviously, the precision of our novel method of 3D gaze-position measurement provided an appropriate basis for the autostereogram study.

The Autostereogram Experiment

Method

Subjects. Eight paid subjects (1 female, 7 male), all of them were students at the University of Bielefeld, participated in this experiment. They had normal or corrected-to-normal visual acuity and were experienced in viewing autostereogram images.

Stimuli. The stimuli in this experiment were autostereogram images whose depth impression arose from two horizontally divided half planes, which were recognized at different virtual distances. The autostereogram images used in the experiment varied in their granularities. Using the autostereogram generation program *rdsgen V1.2b* by Frederic N. Feucht[1], we produced different autostereogram images in which one dot of the autostereogram image corresponded to one, two, four, eight, or twelve screen pixels in height and width. The subjects always got the spatial perception that the lower plane was nearer than the upper one. The depth difference between the two planes corresponded to a vergence angle difference of 0.859^o.

By using *rdsgen* we created two autostereogram images for each of the granularities 1, 2, 4, 8, and 12[2]. These 10 autostereogram images were presented to the subjects in a random order on a 20-inch screen and a spatial resolution of 640x480 pixels. The images consisted of black and blue (luminance: 24 cd/m^2) points, where one pixel corresponded to 4.17 min arc. The distance from the subject to the screen was 50 cm.

---

[1]http://www.bcc.cc.nc.us/graphics/g2c.html.
[2]We first changed the distance between the repeating patterns of the autostereogram image and magnified it with the corresponding "enlargement factor" (grain size) to keep the distance between the repeating patterns (and therefore the depth impression) constant.

Apparatus. We used an SMI EyeLink eye tracker for the experiments. This system employs a headset with two cameras to enable binocular eye-movement recording. Further features of the EyeLink system are a high sampling rate of 250 Hz and an average on-screen gaze position error between 0.5^o and 1.0^o.

Procedure. Prior to the experiment, subjects were presented with autostereogram images without using the eye tracker. This way the subjects could practice the perception of the depth information. A lamp fixed behind the subjects to create reflections on the screen helped the subjects to fixate a virtual point behind the screen plane, facilitating the 3D perception of the autostereogram image. When the subject got the 3D impression, the light from behind was switched off.

At the beginning of every trial the light was switched on as soon as the autostereogram image appeared on the screen. When the subject got the 3D impression of the image, the experimenter switched off the light and started the eye-movement recording. It was the subjects' task to change their view from the front plane to the back plane and vice versa (every few seconds). Between these eye movements subjects kept their view on one plane for a while before changing their view to the other one. The vergence movements were recorded over a duration of 1 minute per image. If subjects "lost" the 3D impression during the recording interval, they had to press the right mouse button, recover the 3D impression, and press the left button to continue recording. Ten autostereogram images were shown to the subject in a random order so that every granularity appeared twice.

Data Analysis. In general, the vergence angle α is used to make statements about a subject's vergence. The bigger the vergence angle is, the stronger the eyes converge, i.e. the nearer the (virtual) surface fixated by the subject. The time course of vergence movements was the most important data to be analyzed in this experiment, especially during the period right before and after subjects changed their gaze point from one plane to another. We distinguished between eye movements from the near to the far plane and vice versa. For the evaluation process we summarized the data of all subjects and calculated the arithmetic mean and the standard error of the vergence angle as a function of time relative to gaze transitions between the two depth planes.

There were two criteria for the presence of a gaze transition between planes: First, there had to be a change in the measured y-value that indicated a crossing of the horizontal boundary. Second, any such change from plane A to plane B had to be preceded by a contiguous sequence of 50 measured gaze positions (200 msec) on plane A, and succeeded by a contiguous sequence of 250 gaze positions (1000 ms) on plane B. To account for a small fraction of measurement errors, we allowed a maximum of 4 violations of these conditions among the 300 measurements. After detecting a plane transition, the time course of vergence from 200 ms before to 1000 ms after the transition was calculated.

Results and Discussion

Our first observation was that subjects had problems to achieve a stable 3D perception of autostereogram images with large grain size (granularities 8 and 12). Once achieved, it was difficult for them to maintain the stable depth impression when changing their gaze point from one plane to the other. Subjects classified the autostereogram images with the granularities 2 and 4 as most pleasant to view. Figure 3 shows the temporal progress of the convergence and divergence movements during the examination of autostereogram images across granularities. The vergence angles are standardized, i.e. 0 corresponds to the back and 1 to the front plane. In the figures the value 0 on the time axis signals the time of plane transition. Each panel shows the vergence data from 200 ms before to 1 s after an eye movement between the two planes. Figure 3 also illustrates that, for all granularities, the convergence movements are obviously faster than the divergence movements. This is in line with the observations in Mowforth, Mayhew and Frisby (1981). The final values are reached after a period of about 800 ms, as stated in Rashbass and Westheimer (1961). As can clearly be seen, both convergence and divergence movements already start before the onset of an eye movement between planes, suggesting that the imagination of the target distance is already sufficient for the execution of vergence eye-movements (proximal vergence). Similar results were already obtained by Yarbus (1967).

The high overshoot of the convergence movements (Figure 3, upper panel) for the granularities 2 and 4 is very obvious. Both show an overshoot of 1.8 and approach the end value at nearly 900 ms. The similarity between those vergence movements is in line with the subjects reporting that it is easier to get and stabilize the 3D illusion of these two types of autostereogram images. One reason could be that the perception of the depth planes is particularly stable for these granularities. Subjects can perceive the near plane very well, even if they fixate a point on the back plane. As a consequence, the visual system might perform a particularly fast, overshooting vergence movement towards an angle that would project the pattern of the near plane to corresponding points on the retina. This mechanism allows us a fast focussing of near or approaching objects, which is an important capability to react in dangerous situations.

The vergence movements during the observation of the autostereogram image with granularity 1 are similar to those for granularities 2 and 4. It is obvious, though, that the overshoot of the convergence movements for granularity 1 is smaller (1.5) than for granularities 2 and 4 (granularity 1 seems too fine for a stable extrafoveal perception). Also the approaching to the end value is shorter (nearly 750 ms instead of approx. 900 ms). These results further support the finding that vergence movements can clearly be driven by relatively high spatial frequencies (fine granularities), as stated in Mowforth, Mayhew and Frisby (1981).

The data for the coarse granularities 8 and 12 are clearly different from those for 1, 2, and 4. Hence it
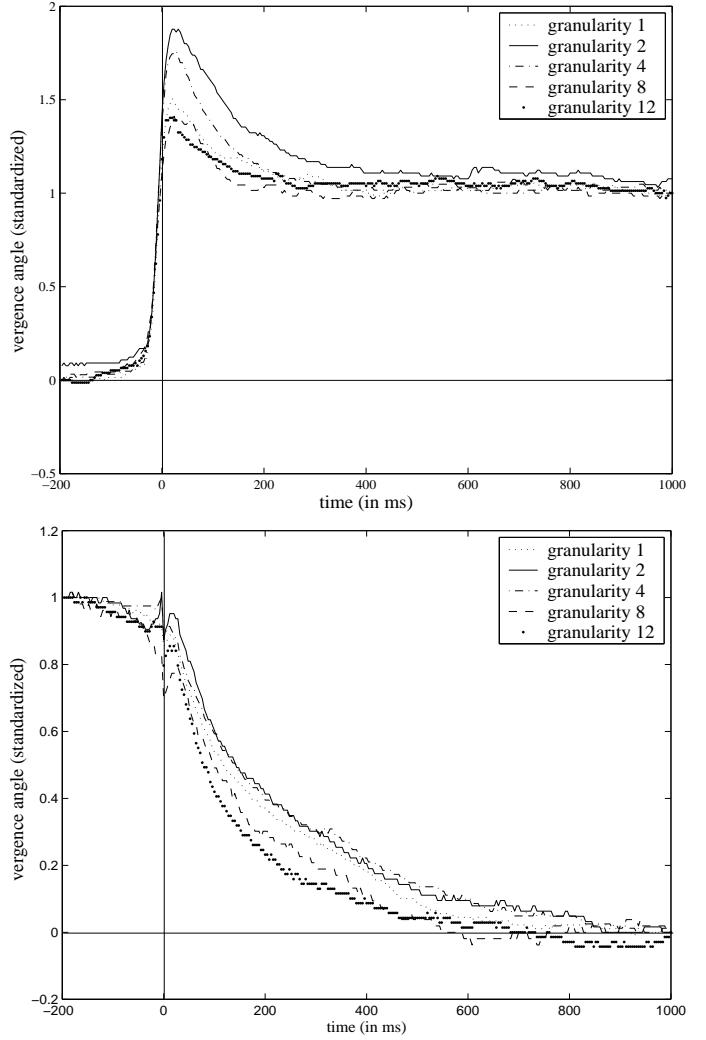

Figure 3: Convergence (upper panel) and divergence movements (lower panel) occurring during the examination of autostereogram images with different granularities.

is much more difficult for the subjects to perceive both planes simultaneously in those autostereogram images, which is consistent with the subjects' impressions. An important point is that the overshoot of the convergence movements is substantially weaker (1.4) than for those occurring in the images with finer granularities. The standard error for the convergence as well as for the divergence movement is high. This can be interpreted as indicating an instable perception of the depth planes. Furthermore, the progress of the divergence movements is faster and finishes sooner than 600 ms after the gaze shift.

The finding of overshoots for convergence movements is inconsistent with the results obtained by Rashbass and Westheimer (1961). It is possible that either the apparatus used in their study did not allow the detection of these overshoots, or that vergence eye-movements in *RDS* – in contrast to those in *SIRDS* – do not show

overshoots. Rashbass and Westheimer describe the vergence system as a damped system, because it does not show any oscillations. In the present study, however, we found overshoots, but no following undershoots or any kind of oscillations. It seems therefore correct to speak of a damped system but with a strong tendency towards overshoots for convergence movements in *SIRDS*.

With regard to divergence movements (Figure 3, lower panel), we found faster approximation of the target angle for the autostereogram images with coarser granularities than for those with finer granularities. This is just the opposite of the results we got for the convergence movements. A possible explanation is a fundamental difference in nature between convergence and divergence movements, as can be seen from Figure 3. Convergence movements are fast and impulse-like, whereas divergence movements are slower and smoother. From an evolutionary standpoint, since distant or vanishing objects are less dangerous, there is no need for fast divergence eye-movement mechanisms. Instead, this divergence resembles more a relaxation process. And this relaxation might be facilitated by easily losing the perception of the near plane. In other words, for the finer granularities 1, 2, and 4, the continuous stable perception of the front plane might impede the subjects' effort to focus attention – and consequently vergence – on the far plane.

All in all, some findings for the *RDS* were confirmed in our experiments with *SIRDS*, in particular that convergence is faster than divergence, that the vergence mechanism can clearly be driven by relatively high spatial frequencies, and that vergence response occurs even before the disparity reaches zero. Inconsistent with previous studies, however, our experiment demonstrates substantial differences in vergence movements between *SIRDS* and previously investigated *RDS*, especially with regard to the overshoot for convergence movements. This difference, however, might have been caused by the possibly low resolution of the apparatus used in the *RDS* study (Rashbass & Westheimer, 1961).

Furthermore, our experiments show a clear influence of granularity on vergence movements. The pattern of influence seems to be more complex than described in Mowforth, Mayhew and Frisby (1981), because, according to our results, lower frequencies (coarser granularities) do not always lead to faster vergence movements. One reason might be the more difficult simultaneous perception of different depth planes in the stimuli of very low frequencies. At any rate, these complex granularity effects on vergence eye-movements in autostereogram images indicate that there may be rather non-intuitive factors that can have a significant impact on the coordination of both eyes. Future research will have to address further issues, such as the timing of coordinated binocular saccades and its dependence on various pattern features. This is a research area that so far has been explored very little, and the rather recent possibility of fast and accurate binocular eye tracking, combined with appropriate techniques for 3D gaze measurement like the approach presented here, will greatly contribute to its advancement.

References

Essig, K. (1998). *Messung von binokularen Augenbewegungen in realen und virtuellen 3D-Szenarien.* Master Thesis, Faculty of Technology, Bielefeld University.

Julesz, B. (1971). *Foundations of Cyclopean Perception.* Chicago: University of Chicago Press.

Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of IEEE, 78,* 1464–1480.

Littmann, E. (1989). Vergenz und Kontrast. *Optometrie, 3,* 15–30.

Logvinenko, A. D. & Belopolskii, V. L. (1994). Convergence as a Cue for Distance. *Perception, 23,* 207–217.

Mowforth, P., Mayhew, J. E. W. & Frisby J. P. (1981). Vergence Eye Movements Made in Response to Spatial-Frequency-Filtered Random-Dot Stereograms. *Perception, 10,* 299–304.

Ning Qian (1997). Binocular Disparity and the Perception of Depth. *Neuron, 18,* 359-368.

Pomplun, M., Velichkovsky, B.M. & Ritter, H. (1994). An artificial neural network for high precision eye movement tracking. In Nebel, B. & Dreschler-Fischer, L. (Eds.), *Lecture notes in artificial intelligence: AI-94 Proceedings,* 63-69. Berlin: Springer Verlag.

Rashbass C. & Westheimer G. (1961). Disjunctive Eye Movements. *J. Physiol., 159,* 339-360.

Reading, R.W. (1983). *Binocular Vision. Foundations and Applications.* Boston: Butterworths.

Ritter, H. (1993). Parametrized Self-Organizing Maps. *ICANN93-Proceedings,* 568–577, Berlin: Springer.

Schor, C. M. & Ciuffreda, K. J. (1983). *Vergence Eye Movements: Basic and Clinical Aspects.* Boston: Butterworths.

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