Pointing in Pictorial Space: Quantifying the Perceived Relative Depth Structure in Mono and Stereo Images of Natural Scenes

MAARTEN W.A. WIJNTJES, SYLVIA C. PONT, Delft University of Technology

Although there has recently been a large increase in commercial 3D applications, relatively little is known about the quantitative perceptual improvement from binocular disparity. In this study we developed a method to measure the perceived relative depth structure of natural scenes. Observers were instructed to adjust the direction of a virtual pointer from one object to another. The pointing data was used to reconstruct the relative logarithmic depths of the objects in pictorial space. The results showed that the relative depth structure is more similar between observers for stereo images than for mono images in two out of three scenes. A similar result was found for the depth range: for the same two scenes the stereo images were perceived as having more depth than the monocular images. In addition, our method allowed us to determine the subjective center of projection. We found that the pointing settings fitted the reconstructed depth best for substantially wider fields of view than the veridical center of projection for both mono and stereo images. The results indicate that the improvement from binocular disparity depends on the scene content: scenes with sufficient monocular information may not profit much from binocular disparity.

Categories and Subject Descriptors: J.4 [**Computer Applicaions**]: Social and Behavioral Sciences—*Psychology*

General Terms: Experimentation, Human Factors, Measurement, Theory

Additional Key Words and Phrases: Depth perception, binocular disparity, natural scenes

ACM Reference Format:

Wijntjes, M. W. A. and Pont, S. C. 2010. Pointing in pictorial space: Quantifying the perceived relative depth structure in mono and stereo images of natural scenes. ACM Trans. Appl. Percept. 7, 4, Article 24 (July 2010), 8 pages.
DOI = 10.1145/1823738.1823742 http://doi.acm.org/10.1145/1823738.1823742

1. INTRODUCTION

Stereopsis has for a long time already been recognized as a visual cue for depth perception [Wheatstone 1838]. Applications such as stereo photography have also been around for decades. Recently, the number of commercial applications based on binocular disparity have increased substantially. This holds for both professional applications such as endoscopic surgery and design-concept visualization as well as consumer applications such as 3D cinema. The effect of binocular disparity on the observer manifests itself both on a high, cognitive level and on a low, perceptual level. At a cognitive level, stereo

This work was supported by a grant from the Netherlands Organisation of Scientific Research (NWO).
Authors' addresses: M. W. A. Wijntjes (corresponding author), S. C. Pont, Delft University of Technology, Industrial Design Engineering, Perceptual Intelligence Lab, Landbergstraat 15, 2628 CE Delft, the Netherlands; email: m.w.a.wijntjes@tudelft.nl.

DOI 10.1145/1823738.1823742 http://doi.acm.org/10.1145/1823738.1823742

information increases the viewing experience and subjective naturalness [Seuntiëns et al. 2005]. The recent success of 3D cinema may partially be due to the cognitive improvement from binocular disparity. On a lower level, the perception of depth should in theory be improved by including disparity. This improvement may be less important for consumer applications but could be crucial for professional applications.

The question how binocular disparity improves 3D perception has been approached via various psychophysical paradigms. This variety of paradigms is necessary to understand the underlying mechanisms of visual perception but it also makes a clear answer rather difficult. Paradigms range from measuring basic geometrical stimulus features such as slant or angular extent up to depth perception in complex scenes. There is an increasing body of literature on the combination of cues for depth perception. These studies often make use of simple stimuli (with good reasons) to investigate which cues are important for depth perception, and how cue combination is achieved. There seems to be general consensus that depth cues, such as texture and disparity [Knill and Saunders 2003; Hillis et al. 2004], combine in a statistically optimal way. Under controlled conditions, binocular disparity plays a significant role in depth perception. When the control is loosened, the role of disparity becomes more difficult to infer from experiments. Pictures of natural scenes contain different combinations of (monocular) depth cues and may in addition be susceptible to top-down influences. Therefore, studies using realistic and complex stimuli are necessary to predict human performance in complex (virtual) environments.

Roughly speaking, the layout of real-world scenes consists of medium boundaries (surfaces), 3D objects with specific shapes, and the spatial organization of these objects with respect to each other. The perception of 3D shape has been studied extensively in a quantitative way, see for example Koenderink et al. [1992, 2001]. The results with respect to the role of binocular disparity are equivocal. For example, disparity was found to be beneficial in a study using random shape stimuli [Norman et al. 1995] but did not improve perception when using human torsos [Todd et al. 1996]. While there is extensive literature on 3D shape perception, there is only little known about the perceived 3D lay-out of a set of objects located in a pictorial space. We introduce a novel paradigm by which the perceived 3D structure of natural scenes can be quantified. Using a pointing task where observers point from one object to another, the relative logarithmic depths can be reconstructed. This paradigm is used to quantify whether stereo information improves visual perception of the 3D structure of natural scenes.

2. METHODS

2.1 Stimuli

For our experiments we needed pictures of scenes that contain distinct objects located at different distances from the camera viewpoint. To meet this requirement we took three pictures: a flock of birds (Figure 1), a bouquet of silk flowers, and a group of people (Figure 3). The birds and people were photographed outdoors and the flower bouquet was photographed in a studio. Pictures were taken with a stereo camera (FinePix REAL 3D W1). This camera has two CCD sensors and two lenses that are parallel to each other, that is, the convergence angle is (approximately) 0 degrees. The distance between the optical axes of the lenses is 77 mm. The outdoor pictures were shot with a horizontal field of view of 54.4 degrees. The indoors picture was shot with a horizontal field of view of 19.5 degrees. We used 11, 13, and 10 pointer/target locations for the birds, flowers, and people scene, respectively.

2.2 Participants

Eight observers volunteered to participate in the experiment. One of them was the first author, another was working in the same lab, and six were students who were reimbursed for their participation. All

Fig. 1. On the left, the birds scene with pointer and target (in the middle at the bottom). Observers were instructed to point from one bird to another. When this is done for all bird pairs, the relative depth structure of the flock of birds can be reconstructed, as illustrated in the middle (frontal view) and right (top view). The average results are indicated by the green and red dots for the mono and stereo condition, respectively.

had normal or corrected-to-normal vision and normal stereo acuity as determined by the TNO stereo-test (acuity all below 60 arc seconds).

2.3 Procedure

Observers were asked to adjust a pointer such that it points towards a target, indicated by a red dot as illustrated in the left image of Figure 1. In each trial, only one pointer and one target were present. The orientation of the pointer could be manipulated with the mouse but was confined to the direction towards the target. In other words, only the slant was adjusted, with a fixed tilt. The pointer and target were only present in the left eye image to avoid (false) disparity information of the pointer and target. The order of scenes was kept constant but the mono/stereo condition was randomized, for example: birds(mono)-flowers(stereo)-people(stereo)-birds(stereo)-flowers(mono)-people(mono). All eight permutations of the mono/stereo condition were used. Participants viewed the stimulus using a Wheatstone stereoscope consisting of two mirrors and two CRT screens. Viewing distance was 68 cm and stimuli were displayed full screen (40×30 cm).

2.4 Data Analysis

2.4.1 *Depth from Depth Differences.* First we will show how to reconstruct (relative) depths from a set of depth differences. For a set of n objects, $\{z_n\}$, $n(n-1)/2$ depth differences $\{z_i - z_j\}$ exist. Thus we can formulate a linear equation that relates the depth vector $\mathbf{z} = \{z_1, z_2, \ldots\}$ of length n with a depth difference vector $\Delta\mathbf{z}$ of length $n(n-1)/2$ as

$$\mathbf{M}\mathbf{z} = \Delta\mathbf{z}, \tag{1}$$

where $\mathbf{M}$ is a $n(n-1)/2$-by-n matrix. The elements of Eq. (1) are:

$$\begin{pmatrix} 1 & -1 & 0 & 0 & \ldots \\ 1 & 0 & -1 & 0 & \ldots \\ 1 & 0 & 0 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ldots \\ 0 & 1 & -1 & 0 & \ldots \\ 0 & 1 & 0 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ \vdots \end{pmatrix} = \begin{pmatrix} z_1 - z_2 \\ z_1 - z_3 \\ z_1 - z_4 \\ \vdots \\ z_2 - z_3 \\ z_2 - z_4 \\ \vdots \end{pmatrix}. \tag{2}$$

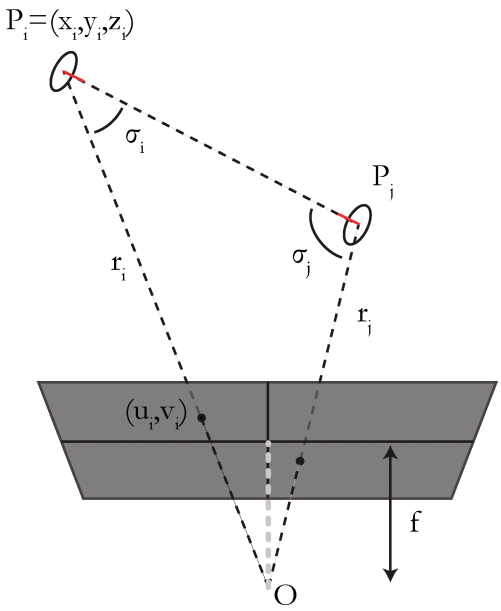

Fig. 2. Illustration of the reconstruction method.

This system of linear equations is overdetermined, hence we need the pseudo inverse matrix $\tilde{\mathbf{M}}^{-1}$ to compute the least squares solution:

$$\tilde{\mathbf{z}} = \tilde{\mathbf{M}}^{-1}\Delta\mathbf{z}\,. \tag{3}$$

It is likely that the pointing settings are not completely coherent, that is, that Eq. (1) cannot be fulfilled. The amount of incoherence (ϵ) in the data can be quantified by taking the norm of the difference between the actual settings $\Delta\mathbf{z}$ and the reconstructed settings $\mathbf{M}\tilde{\mathbf{z}}$:

$$\epsilon = |\mathbf{M}\tilde{\mathbf{z}} - \Delta\mathbf{z}|\,. \tag{4}$$

2.4.2 *Measurement of Depth Differences.* Using a pointing paradigm, it is possible to measure the depth difference between objects. This would be straightforward if the projection were orthographic. Since this is generally not the case, and certainly not for our stimuli, we used a method that does not depend on the projection, as long as it is linear. Two objects (P_i and P_j) and the observer (O) together constitute a triangle OP_iP_j (Figure 2). The length of the triangle legs $|OP_i| = r_i$ and $|OP_j| = r_j$ can be computed using the sine rule:

$$\frac{\sin\sigma_i}{r_j} = \frac{\sin\sigma_j}{r_i}\,. \tag{5}$$

Since a linear equation is needed for the depth reconstruction algorithm, we use the logarithmic depths:

$$\log r_i - \log r_j = \log\left(\frac{\sin\sigma_j}{\sin\sigma_i}\right). \tag{6}$$

Thus, instead of the depth values z_i, we reconstruct the logarithmic distances $\log r_i$.

A reconstruction based on Eq. (3) evidently gives the relative depth, that is, plus or minus some constant factor, $z \to z + a$. In the logarithmic representation this means an arbitrary scaling factor, $r \to ar$.

2.4.3 *Reconstruction of the 3D Layout.* To reconstruct the 3D lay-out of all objects $\{P_i\}$ from the subjective (logarithmic) distances r_i, we need the image or screen coordinates (u, v) and the focal length f. It can readily be seen from Figure (2) that the 3D coordinates can be written as:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \frac{r_i}{\sqrt{u_i^2 + v_i^2 + f^2}} \begin{pmatrix} u_i \\ v_i \\ f \end{pmatrix}. \tag{7}$$

The interesting aspect of this formula is not so much that it allows to plot the data in 3D space, but that we can analyze what the subjective center of projection was that observers used when viewing the scenes. It is well known that observers can view and perceive images when looking from other points than the center of projection. Our images were not shot with a single focal length and the observers were not positioned at the proper viewing angle. In general, users of stereo applications, such as cinema visitors, will also not be properly positioned and the cameras will zoom in and out while the screen size stays the same. With Eq. (7) we can use the focal length f as a fit parameter and analyze the subjective focal position taken by the observer. Note that the scaling factor is irrelevant since it applies both to the 3D coordinates (x_i, y_i, z_i) and the distance r_i. The orientations of the pointers can be calculated from the points P_i using the cosine rule:

$$\tilde{\sigma}_i(f) = \cos^{-1}\left(\frac{r_i^2 + |\mathbf{P}_i - \mathbf{P}_j|^2 - r_j^2}{2r_i|\mathbf{P}_i - \mathbf{P}_j|}\right). \tag{8}$$

Because the points P_i are a function of f, so is the reconstructed orientation $\tilde{\sigma}_i$. The cost function that is minimized to find the subjective focal point (viewing angle) is the norm of all reconstructed orientations $\tilde{\sigma}_i$ and experimental orientations σ_i.

2.4.4 *Statistical Analysis.* Each scene was tested with a (paired or unpaired, depending on the type of variable) t-test, Bonferroni corrected for the three comparisons. Significance level was set at $\alpha = 0.05$. Since we were not interested in absolute differences between the scenes, we only statistically analyzed differences within each scene.

3. RESULTS

An illustration of the average 3D reconstructions can be seen in Figures 1 and 3. First, the inconsistency of the settings as defined by Eq. (4) was analyzed, which can be seen in Figure 4(a). The bars denote the average results and the individual data are illustrated by the connected dots. The inconsistency of the settings seems generally higher for stereo images. However, this could not be confirmed statistically for any of the scenes.

Second, we analyzed the reconstructed depths. To assess the similarity of depth perception within a scene, we correlated the depth values of the different observers with each other. This can be regarded as a measure for the consistency of depth perception across observers, rather than the internal consistency of the pointing settings as analyzed before. The results are shown in Figure 4(b). Depth perception is more consistent between observers for the stereo images of the bird and flower scene, but not for the people scene. We also calculated the correlation within observers, between the mono and stereo conditions, which is illustrated by the dotted grey line. For the bird and people scenes this correlation seems of comparable magnitude as the within mono results.

Third, the depth range was quantified by the standard deviation of the log depths. For the birds and flowers scenes, the stereo images elicit more depth than the mono images, but not for the people scene (Figure 4(c)).

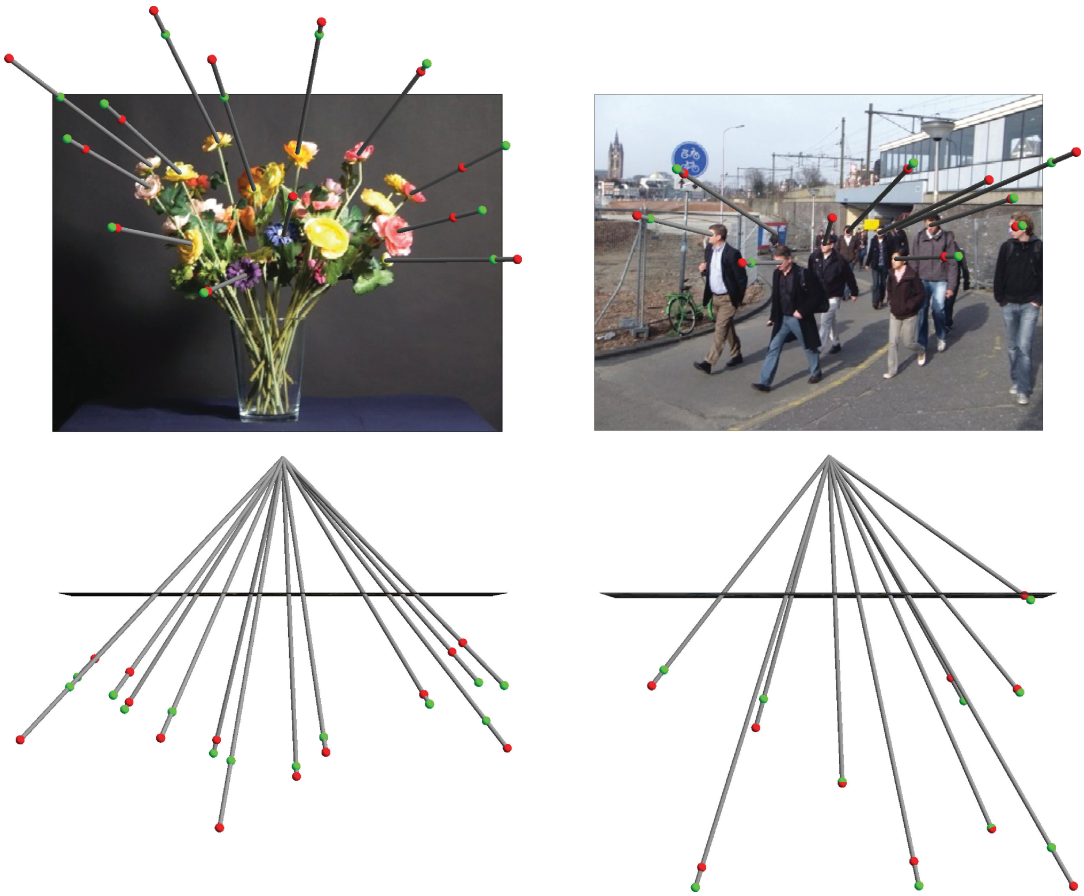

Fig. 3. Front and top view of the average depth reconstructions for the flower and people stimulus. Red indicates the stereo condition, green the mono condition.

Finally, we analyzed the subjective viewing angle. Results are shown in Figure 4(d) together with the experimental viewing angle (red dashed line) and the camera angles (blue dashed lines). The only statistical difference between the conditions was found for the flowers scene. Furthermore, subjective viewing angles were always larger than both the experimental and camera angles.

4. DISCUSSION

We used a novel method to quantify the perceived depth structure of an image. In this article we have focused on the added value of binocular disparity in the context of 3D applications. However, the method is generic and can be used to approach various other questions related to pictorial space. Both the task and the reconstruction algorithm are relatively easy to program. Furthermore, observers reported that the task was intuitive and easy to use.

There seems to be a counterintuitive tendency of higher inconsistency for stereo images, although the results were not significant. For all three scenes, 6 out of 8 observers showed this effect. One would expect that in a scene where more depth information is available, settings would be more consistent (i.e., less inconsistent). Possibly, the perceptual space for a monocularly viewed scene is more consistent

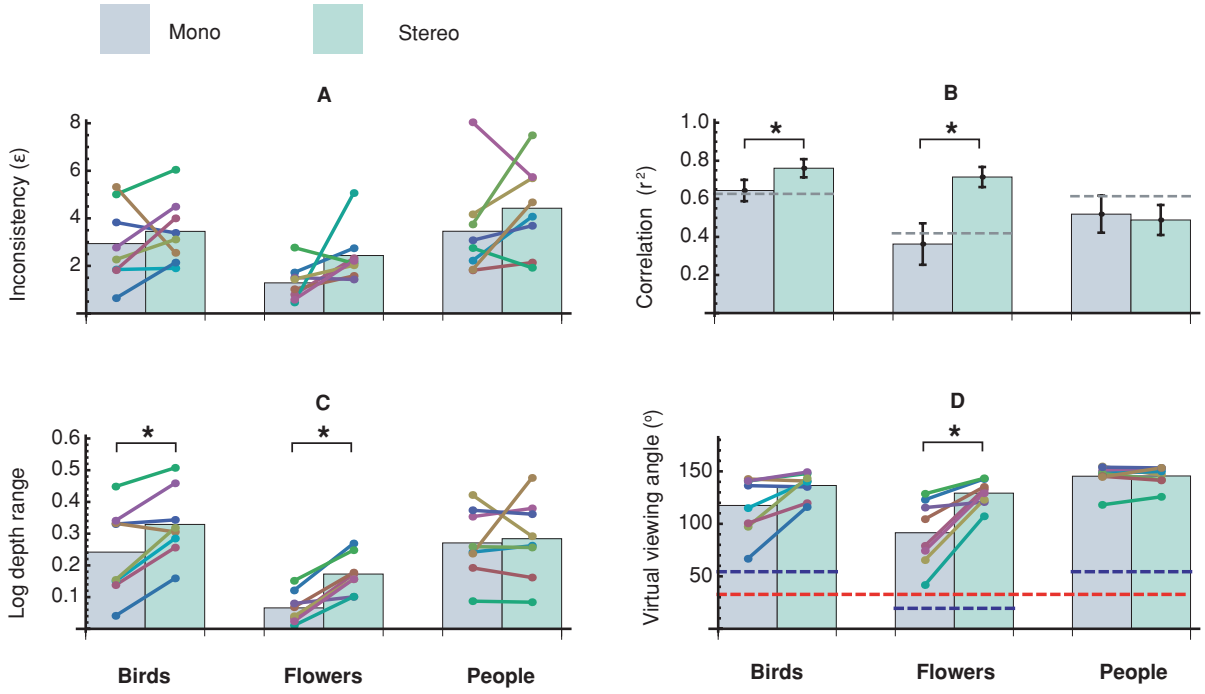

Fig. 4. Results of the experiment for each scene and mono-stereo condition. (a) The inconsistency as defined by Eq. (4), lines denote results of individual subjects. (b) Mean correlation (95% confidence intervals) between the depth values of all observer pairs within each condition. (c) Standard deviation of the log depths per scene. (d) Virtual viewing angle.

because in the stereo image, disparity conflicts with vergence and accommodation [Hoffman et al. 2008]. Another explanation could be that due to the larger depth range results, the noise of the settings is also larger which could lead to a higher inconsistency.

Although less internally consistent, the subjective depth structures are more similar between observers for stereo images than for mono images, at least for the flowers and birds scenes. The fact that this effect is not present in the people scene might be due to the fact that monocular cues are stronger in that scene. In the birds scene, the objects float in the void, and only relative size differences contribute to monocular depth perception. In case of the flowers scene, the objects are in some way connected to each other but the bouquet may be too cluttered to make sense out of these connections. On the other hand, the objects in the people scene are all located on the ground and can therefore be monocularly perceived. These different results for different scene content can be used in various applications because it shows that stereo information may be more beneficial in scenes where monocular cues are relatively weak. Besides a more similar depth structure, the stereo images also elicit higher depth ranges, a result that is in line with previous findings in 3D shape perception [Koenderink et al. 1995].

The subjective centers of projection seem to be rather puzzling. In general, the subjective viewing angles are much larger than either the camera or the experimental viewing angles. This is true for both mono and stereo images. This effect can be due to systematic errors in slant settings or it could be that observers actually use a “mental viewpoint” [Koenderink et al. 2001] that deviates considerably from both the focal length and viewing distance. The possibility of systematic errors in slant settings can be understood by looking at Figure 2: if the observer location O moves towards the screen (along the

grey dashed line), the viewing angle increases. This means that the other two angles should decrease. Hence, if the slant settings are systematically underestimated, the subjective center of projection will decrease (move towards the screen) resulting in an increased subjective viewing angle. This can be further investigated by comparing the pointing method from our current research with a different method (such as pair-wise depth discrimination) that does not make use of the pointing probe. If we still find large subjective viewing angles without using a pointing probe, the effect does not depend on systematic slant biases. In that case, the effect could be due to an actual shift of mental viewpoint, which can be investigated by systematically manipulating focal length and viewing distance.

REFERENCES

HILLIS, J. M., WATT, S. J., LANDY, M. S., AND BANKS, M. S. 2004. Slant from texture and disparity cues: Optimal cue combination. *J. Vis. 4*, 12, 967–992.

HOFFMAN, D., GIRSHICK, A., AKELEY, K., AND BANKS, M. 2008. Vergence-Accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vis. 8*, 3.

KNILL, D. AND SAUNDERS, J. 2003. Do humans optimally integrate stereo and texture information for judgements of surface slant? *Vis. Res. 43*, 2539–2558.

KOENDERINK, J. J., VAN DOORN, A. J., AND KAPPERS, A. M. 1995. Depth relief. *Percep. 24*, 1, 115–126.

KOENDERINK, J. J., VAN DOORN, A. J., AND KAPPERS, A. M. L. 1992. Surface perception in pictures. *Percep. Psychophys. 52*, 487–496.

KOENDERINK, J. J., VAN DOORN, A. J., KAPPERS, A. M. L., AND TODD, J. T. 2001. Ambiguity and the 'mental eye' in pictorial relief. *Percep. 30*, 4, 431–448.

NORMAN, J. F., TODD, J. T., AND PHILLIPS, F. 1995. The perception of surface orientation from multiple sources of optical information. *Percep. Psychophys. 57*, 5, 629–636.

SEUNTIËNS, P. J. H., HEYNDERICKX, I. E. J., IJSSELSTEIJN, W. A., VAN DEN AVOORT, P. M. J., BERENTSEN, J., DALM, I. J., LAMBOOIJ, M. T. M., AND OOSTING, W. 2005. Viewing experience and naturalness of 3d images. In *Proc. SPIE - The Inter. Soc. Optical Engin.* 6016, 601605.

TODD, J. T., KOENDERINK, J. J., VAN DOORN, A. J., AND KAPPERS, A. M. L. 1996. Effects of changing viewing conditions on the perceived structure of smoothly curved surfaces. *J. Exper. Psychol.-Hum. Percep. Perform. 22*, 3, 695–706.

WHEATSTONE, C. 1838. Contributions to the physiology of vision–Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosoph. Trans. Roy. Soc. London 128*, 371–394.

Received June 2010; accepted June 2010