Top-down influences on stereoscopic depth-perception

Isabelle Bülthoff[1], Heinrich Bülthoff[1] and Pawan Sinha[2]

[1] *Max-Planck-Institut für biologische Kybernetik, 72076 Tübingen, Germany*

[2] *Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706, USA*

Correspondence should be addressed to I.B. (isabelle.buelthoff@tuebingen.mpg.de)

The interaction between depth perception and object recognition has important implications for the nature of mental object representations and models of hierarchical organization of visual processing. It is often believed that the computation of depth influences subsequent high-level object recognition processes, and that depth processing is an early vision task that is largely immune to 'top-down' object-specific influences, such as object recognition. Here we present experimental evidence that challenges both these assumptions in the specific context of stereoscopic depth-perception. We have found that observers' recognition of familiar dynamic three-dimensional (3D) objects is unaffected even when the objects' depth structure is scrambled, as long as their two-dimensional (2D) projections are unchanged. Furthermore, the observers seem perceptually unaware of the depth anomalies introduced by scrambling. We attribute the latter result to a top-down recognition-based influence whereby expectations about a familiar object's 3D structure override the true stereoscopic information.

Our visual system can often recognize objects not only on the basis of their static appearance but also by observing how they move. In some cases, impoverished image sequences can be recognized from their pattern of motion despite the fact that no single frame has enough figural information to support recognition (Fig. 1). Image sequences such as those devised by Johansson[1], which convey vivid impressions of humans engaged in various dynamic activities, are elegant and powerful demonstrations of this fact.

Since the first reports of this work, the issue of how motion sequences are interpreted by the primate brain has been extensively studied. Previous work[2] has shown that observers can identify human individuals from their gait alone. Coding theory principles have also been applied to model gait perception[3]. A model for the interpretation of these sequences has been developed[4] by adapting structure-from-motion ideas[5]. Similar approaches have also been developed by other researchers [6,7]. Neurons that respond selectively to biological motion sequences have been described in the visually responsive areas of the primate temporal cortex[8], a region that is known to be involved in object recognition.

Most of the studies that have attempted to explain human perception of biological motion sequences have done so largely in terms of bottom-up mechanisms, in which the object geometry is extracted from low-level features without recourse to higher-level internal representations of objects. In this report, however, we examine the possibility that internal object representations may also play a top-down role in this process. Our stimuli were stereo views of walking human figures that were defined by a small number of dots; we have used depth-distorted versions of these figures to study the interactions between depth cues and recognition. Our results provide two pieces of evidence in this regard. First, they suggest that the anomalous stereo-depth cues do not significantly influence the recognizability of the stimuli. Second, they show that top-down recognition-based influences can strongly alter depth perception, such that expectations about a familiar object's 3D structure override the true stereoscopic information. Consistent with this hypothesis, we have found that reducing the recognizability of objects reduces the magnitude of the top-down influence on depth perception.

Results

As stimuli, we used variants of previously described biological motion sequences[1] (see Fig. 1). They were 3D stereo animations showing twelve points on a male human (three points positioned at the joints of each limb) as he walked on a treadmill at a normal pace. We compared the normal version of this stimulus with a 'depth-scrambled' version in which the depth positions of the joints were randomly altered in the *z*-axis. This rendered their 3D trajectories arbitrary within the volume defined by the original structure, while leaving their 2D projections on the retina unchanged (discounting the minor positional shifts induced by the need to incorporate depth-disparity information). The degree of arbitrary depth scrambling of the scrambled walker was a continuously variable parameter. We defined the extent of added depth noise as a function of the depth-extent (say, D) of the original undistorted walker. Thus, a noise level of 0 % corresponded to an undistorted sequence, whereas a noise level of 50 % implied a sequence wherein the individual body points could assume any depth value (with uniform probability) within a bound of D/2 about their original depth position. Additionally, we generated 'random' versions of both the normal and the depth-scrambled walkers by scrambling the *x* and *y* positions of their constituent points. Unlike the purely depth-scrambled version, this created distortions of the 2D retinal projection. In all sequences, the added offsets were kept constant from frame to frame.

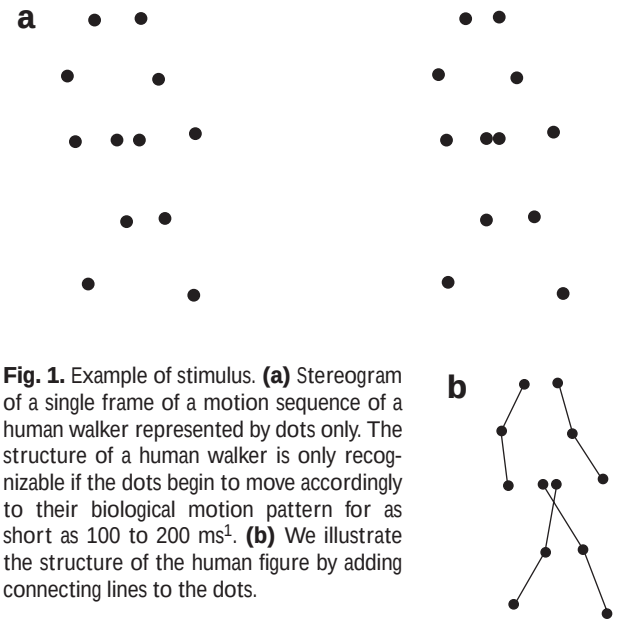

Fig. 1. Example of stimulus. **(a)** Stereogram of a single frame of a motion sequence of a human walker represented by dots only. The structure of a human walker is only recognizable if the dots begin to move accordingly to their biological motion pattern for as short as 100 to 200 ms[1]. **(b)** We illustrate the structure of the human figure by adding connecting lines to the dots.

In our first experiment (recognizability experiment), we asked whether randomizing the depth structure of the moving figure while preserving its 2D traces would adversely affect its recognizability as a human. Subjects viewed in stereo the distorted walker sequences interspersed with random and human (unscrambled) sequences. The viewing position was varied between 0° (the figure is seen walking in place facing right; depth distortion does not affect the positions of the points in the image plane) and 90° (the walker is walking toward the observer; the same distortion now results in displaced points in the new image plane, see Fig. 2). The subjects rated all sequences for their structural goodness as a human on a scale from 1 (completely random) to 5 (completely human). We expected that if stereo-depth information were critical for the recognition processes, subjects would perceive our depth-scrambled sequences as random objects from all viewpoints because the 3D structure of the depth-distorted walker would be completely different from a human figure. The ratings assigned to these sequences would, therefore, be uniformly low for all viewing positions. If, however, recognition required merely 2D congruence, then the rating would be expected to increase in going from a viewing direction perpendicular to the distorted depth axis (90°) to one parallel to it (0°), as the monocular view comes to resemble more closely a 2D human figure.

The key result (Fig. 3) is that irrespective of the depth structure of the sequences, viewpoints preserving the 'normal' 2D projections yielded biological motion percepts (high ratings). Other viewpoints for the scrambled sequence yielded percepts of randomly moving dots (low ratings). The data strongly indicate that the recognition process used by the subjects in this task is heavily biased towards 2D traces. Stereo-depth information does not seem to contribute significantly to the recognition processes.

The most surprising result of this experiment is that depth-scrambled motion sequences that had 'normal' 2D traces were rated as highly as unscrambled sequences. There are at least two possible explanations for this. Either subjects might be perceptually aware of the depth scrambling but decide nevertheless to base their ratings on the similarity of the 2D projection, or they might be perceptually unaware of the depth scrambling, possibly due to a top-down object-specific influence that actively imposes the expected structure on the input and thus suppresses the perception of 3D anomalies.

To distinguish between these possibilities, we designed a second experiment (depth-plane experiment) to test for the existence of any recognition-dependent influences that might serve to suppress information about depth-anomalies being provided by low-level stereo processes. To assess observers' ability to perceive the true depth structure of these sequences, we designed a simple task that required them to report whether three indicated points in the structure were in the same fronto-parallel depth plane. Stereo viewing was used throughout the experiment. Each experimental trial commenced with a presentation of either a depth-scrambled walker or a random pattern for the duration of one walk-cycle, this being sufficient time to allow the moving figure to be recognized. As the presentation continued into the next cycle, three of the points were highlighted by thin red outlines. After two-thirds of the duration of a walk-cycle, the screen then turned blank. Subjects had been instructed to report whether the three red dots lay in the same fronto-parallel depth plane. In 50 % of the sequences, the three dots were in the same plane and in 50% they were not. In the depth-distorted figures, we could vary the depth of the highlighted dots independent of their positions on the limbs (same versus different limb); this allowed us to ask whether the perceived depth was influenced by the expectation that points on the same limb would be at the same depth.

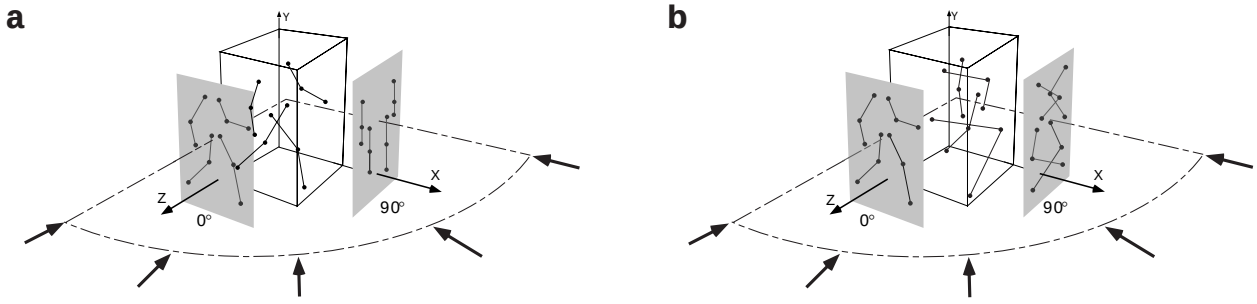

Fig. 2. In the recognizability experiment, subjects viewed depth-scrambled and normal sequences in stereo from different viewing positions indicated by the arrows along the equator at waist level in (a) and (b). **(a)** Undistorted walker. **(b)** Depth-scrambling a biological motion sequence involves adding depth noise to the positions of the joints while leaving their 2D positions (in the *xy*-plane) largely unchanged. From other viewing positions (e.g., in the *yz* projection shown on the right), the original 2D pattern of a human figure is severely distorted. In all figures, connecting lines serve to enhance recognizability of the human figure and are not shown in the experiments.

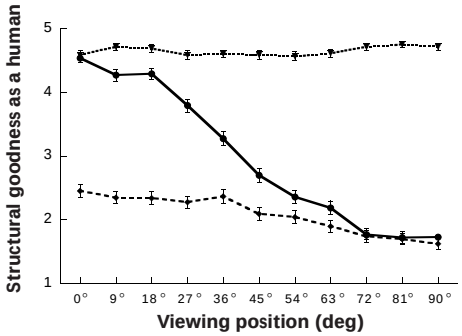

Fig. 3. Results of the recognizability experiment averaged across 22 subjects. The following stimulus sequences were presented: human walker (no distortion), depth-distorted walker (a constant *z*-distortion was applied from frame to frame, the amount of depth noise added was 100%) and random pattern (constant *xz*-distortion). The subjects rated them on a scale from 1 (very random) to 5 (very human). Abscissa: viewing position in degrees, at 0° the walker is seen walking to the right with its depth axis parallel to the viewing axis; at 90° the walker is seen walking towards the observer with its depth axis perpendicular to the viewing axis. P *z*-distortion, R *xz*-distortion, H no distortion. $n = 22$.

In Fig. 4a and b, the false-alarm rate (a response of 'in same plane' when the dots are in fact not in the same plane) is shown as a function of depth-noise level and depth-disparity respectively. The curves show data for three stimulus types: (1) human sequences with dots on the same limb, (2) human sequences with dots on different limbs, and (3) random sequences. If the perceived depth is affected by prior expectations based on object recognition, the dots on the same limb would be most likely to be perceived as coplanar, and those on different limbs would be least likely. The false alarm rate is highest for the 'same limb' condition. In Fig. 4c, the hit-rate (correct responses of 'in same plane') is shown as a function of depth noise. It is uniformly high for the 'same limb' condition (average value, 94%, SE 1%), lower for the random sequences (average value, 72%, SE 1%) and lowest for the 'different limbs' condition. Thus, the perceived depth is influenced by prior expectations about the depth structure of the image, which are in turn determined by its recognizability.

Discussion

Our results have important implication for the nature of the mental representations for dynamic objects. The limited influence of depth information on object recognition observed in the first experiment suggests that the recognition process is based largely on matching the stimulus to an internal representation of the object's 2D trace-structure rather than its 3D geometry. Consistent with this idea, the monkey infero-temporal cortex, which is known to be involved in object recognition, has recently been reported to contain 'view-tuned' neurons, which respond to 3D objects only when they are seen from a certain viewpoint[9]. An alternative possibility, which we must consider, is that recognition might involve structure-from- motion processes. This idea is based on the fact that a vivid perception of 3D structure can arise from the 2D projection of a rotating object, in the absence of stereoscopic depth cues. Structure-from-motion perception can occur independent of object recognition, because even unfamiliar rotating objects give rise to a 3D percept. Our subjects might therefore have derived a 3D structure from the moving 2D projection of the walking figure and matched this to an internal 3D representation. We believe, however, that this is unlikely, because it has been[10] demonstrated that recognition based on 2D cues proceeds unhindered even when the 3D structure suggested by structure-from-motion processes is inconsistent with the object identity.

We interpret the results of the depth-plane experiment as pointing to the existence of a top-down influence capable of modulating the information provided by the early depth-perception processes based on binocular disparities. There are other related examples, e.g. the hollow mask effect[11] and the cyclopean Necker-cube[12] (in which an ambiguous 2D image gives rise to two alternating percepts with different depth structures in conflict with the disparity given by the stereogram), which argue in favor of the influence of high-level cues on depth perception. This influence can, in turn, be

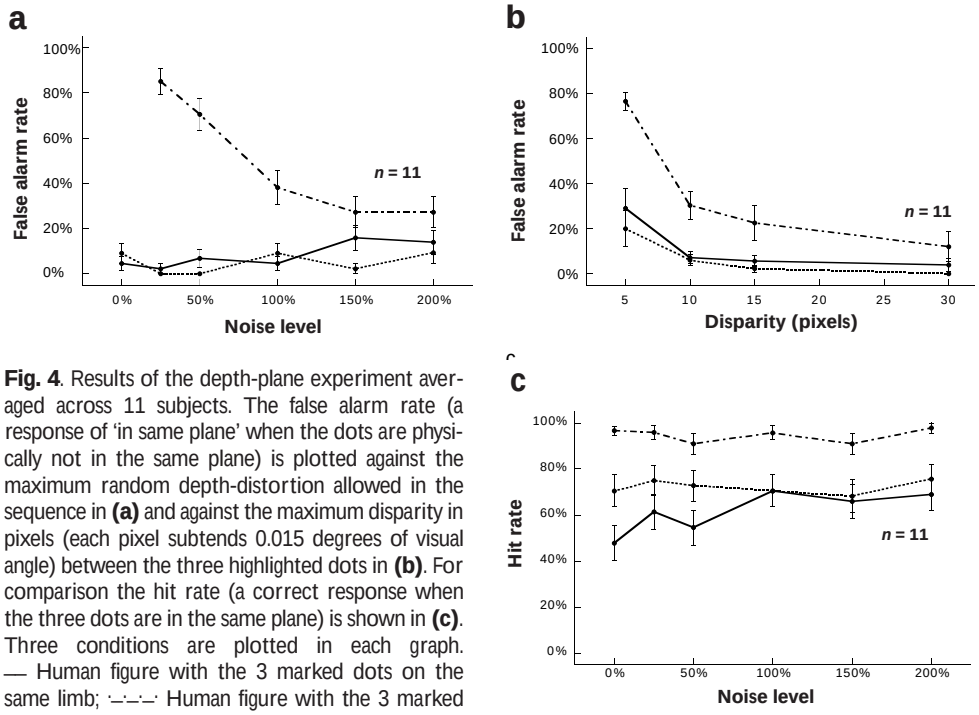

Fig. 4. Results of the depth-plane experiment averaged across 11 subjects. The false alarm rate (a response of 'in same plane' when the dots are physically not in the same plane) is plotted against the maximum random depth-distortion allowed in the sequence in **(a)** and against the maximum disparity in pixels (each pixel subtends 0.015 degrees of visual angle) between the three highlighted dots in **(b)**. For comparison the hit rate (a correct response when the three dots are in the same plane) is shown in **(c)**. Three conditions are plotted in each graph. — Human figure with the 3 marked dots on the same limb; ·–·–·–· Human figure with the 3 marked dots on different limbs; ····· Random figure. The viewing position for all trials was 0° (the 'walker' is seen walking to the right); this viewing position insured good recognizability of the moving figure. The specific noise levels we used for the distorted walker in this experiment were 0, 25, 50, 100, 150, and 200 %, which is around the noise level used in the first experiment.

modulated by factors that change the recognizability of a stimulus. This hypothesis would explain why the human sequences give rise to false depth perception so much more frequently than do the random sequences. Other important factors that need to be considered in interpreting these results include grouping induced by common motion or proximity. That is, dots may be more likely to be perceived as coplanar if they move in synchrony or if they are close together. We believe, however, that in our experiments the effect of such factors would be relatively limited, because the sequences in the different conditions had very similar mid-level attributes such as motion and density distributions. Specifically the differences in performance between the walker and the random conditions suggest the importance of object-specific, recognition-based influences over general configurational ones. The motion trajectories of the individual dots were the same in both conditions, and only their 2D offsets were randomized. Thus, two dots of the walker that moved in phase continued to do so in the random stimulus, thereby largely maintaining the mid-level grouping cues and the articulation geometry in the two conditions. Yet, the suppression of binocular disparity perception occurs only when the walker is recognizable.

The idea that top-down influences can affect perception is certainly not a new one. Several well known visual illusions, such as the Dalmatian dog picture[13] or the mother-in-law/daughter-in-law figure[14] demonstrate the significance of top-down expectations in interpreting ambiguous stimuli. Recent computational models of the neocortex have argued that feedback cortico-cortical projections might allow top-down influences to propagate from the higher cortical areas that are involved in object recognition back to the earlier areas that support lower-level processes[15,16]. Our study now provides evidence that even the very low-level process of stereo-depth perception, which was previously considered to be a purely bottom-up process[17,18], is in fact susceptible to top-down influences. Additionally, our experimental results provide indirect evidence that dynamic three-dimensional objects might be recognized by the visual system based on their 2D traces rather than on their 3D structural descriptions.

Methods

The biological motion sequences used in our experiments were based on data collected at the Gait Analysis Laboratory of the Spaulding Rehabilitation Hospital in Boston, Massachusetts. The data comprised the 3D positions of twelve points on a male human as he walked in place. The point positions were updated 39 times over the course of one complete walk cycle. The stimuli were generated by displaying each point as a bright dot (0.03 degree of visual angle) on a gray background (mean luminance: 20 cd per m^2). The experiments were conducted on a Silicon Graphics Indigo 2 workstation. All sequences were presented in stereo using a pair of StereoGraphics Crystal-Eyes (TM) LCD shutter glasses synchronized with the display. Two views were generated for each frame to allow stereoscopic vision. All subjects were tested to ensure that they had functioning stereoscopic ability, two subjects were rejected and all subjects were naive as to the purpose of the experiments.

Acknowledgments

We wish to thank N. Logothetis, B. Tjan and D. Kersten for insightful comments on earlier versions of the manuscript and P. Lipson for providing the biological motion data set.

RECEIVED 29 JANUARY; ACCEPTED 22 MAY 1998

1. Johansson, G. Visual perception of biological motion and a model of its analysis. *Percept. Psychophys.* 14, 201–211 (1973).
2. Cutting, J. E. & Kozlowski, L. T. Recognition of friends by their walk. *Bull. Psychonom. Soc.* 9, 353–356 (1977).
3. Cutting, J. E. Coding theory adapted to gait perception. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 71–87 (1981).
4. Hoffman D. D. & Flinchbaugh B. E. The interpretation of biological motion. *Biol. Cybern.* 42, 195–204 (1982).
5. Ullman S. *The Interpretation of Visual Motion* (MIT Press, Cambridge, 1979).
6. Shibata T., Sugihara, K., & Sugie, N. Recovering three-dimensional structure and motion of jointed objects from orthographically projected optical flow. *Trans. IECE* 68-D, 1689–1696 (1985).
7. Webb J. & Aggarwal, J. Structure from motion of rigid and jointed objects. *Artif. Intell.* 19, 107–131 (1982).
8. Oram, M. W. & Perrett, D. I. Responses of anterior superior temporal polysensory (STPa) neurons to 'biological motion' stimuli. *J. Cog. Neurosci.* 6, 99–116 (1994).
9. Logothetis, N. K., Pauls, J. & Poggio, T. Shape recognition in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563 (1995).
10. Sinha, P. & Poggio, T. Role of learning in three-dimensional form perception. *Nature* 384, 460–463 (1996).
11. Gregory, R. L. in *Illusion in Nature and Art* (eds Gregory, R. L. & Gombrich, E. H.) 49–96 (Duckworth, London, 1973).
12. Julesz, B. *Foundations of Cyclopean Perception* (Univ. of Chicago, Chicago, 1971).
13. Goldstein, B. *Sensation and Perception* (Brooks/Cole, Pacific Grove, 1996).
14. Boring, E. G. A new ambiguous figure. *Am. J. Psychol.* 42, 444–445 (1930).
15. Ullman, S. Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cereb. Cortex* 5, 1–11 (1995).
16. Mumford, D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* 66, 241–252 (1992).
17. Julesz, B. Early vision and focal attention. *Rev. Mod. Phys.* 63, 735–772 (1991).
18. Nakayama, K., Shimojo, S. & Silverman, G. H. Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* 18, 55–68 (1989).