The Perception of Spatial Layout in a Virtual World

Heinrich H. Bülthoff[1] and Chris G. Christou[2]

[1] Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany
heinrich.buelthoff@tuebingen.mpg.de
[2] Unilever Research, Wirral, UK.

Abstract. The perception and recognition of spatial layout of objects within a three-dimensional setting was studied using a virtual reality (VR) simulation. The subjects' task was to detect the movement of one of several objects across the surface of a tabletop after a retention interval during which time all objects were occluded from view. Previous experiments have contrasted performance in this task after rotations of the observers' observation point with rotations of just the objects themselves. They found that subjects who walk or move to new observation points perform better than those whose observation point remains constant. This superior performance by mobile observers has been attributed to the influence of non-visual information derived from the proprioceptive or vestibular systems. Our experimental results show that purely visual information derived from simulated movement can also improve subjects' performance, although the performance differences manifested themselves primarily in improved response times rather than accuracy of the responses themselves.

1 Introduction

As we move around a spatial environment we appear to be able to remember the locations of objects even if during intervening periods we have no conscious awareness of these objects. We are for instance able to remember the spatial layout of objects in a scene after movements and predict where objects should be. This ability requires the use of a spatial representation of the environment and our own position within it. Recent experiments have shown that although people can perform such tasks their performance is limited by their actual experience of the scene. For instance, Shelton & McNamara [11] found that subjects ability to make relative direction judgements from positions aligned with the studied views of a collection of objects were superior to similar judgements made from misaligned positions. In general it has been shown that accounting for misalignments in view requires more effort in as much as response times and error rates are higher than for aligned views. Similar findings were also reported by Diwadkar & McNamara [5] in experiments requiring subjects to judge whether a configuration of several objects was the same as a configuration of the same objects studied previously from a different view. They found that response latencies

S.-W. Lee, H.H. Buelthoff, T. Poggio (Eds.): BMCV 2000, LNCS 1811, pp. 10–19, 2000.

were a linear function of the angular distance between the test and previously studied views.

These findings and others (e.g., see [9]) have led researchers to conclude that the mental representation of spatial layout is egocentric in nature and encodes the locations of objects with respect to an observer-centred reference frame. However, Simons & Wang [12] found that the manner in which the transformation in view is brought about is important. For instance, people can compensate better for changes in the retinal projection of several objects if these changes are brought about by their own movement. These experiments contrasted two groups of subjects. The first group performed the task when the retinal projection of objects was a result of the (occluded) rotation of the objects themselves while for the second group it was a result of their own movement. In both cases subjects performed equivalent tasks; that is, to name the object that moved when the retinal projection was both the same and different compared to an initial presentation of the objects. Simons & Wang attributed the superior performance by the 'displaced' observers to the involvement of extra-retinal information which, for instance, may be derived from vestibular or proprioceptive inputs. Such information could allow people to continually update their position in space relative to the configuration of test objects. This is known as spatial updating.

This result supports the findings in other spatial layout experiments which contrasted imagined changes in orientation with real yet blind-folded changes in orientation [10,8,6]. The subjects' task was to point to the relative locations of objects from novel positions in space after real or imagined rotations or translations. It was found that translation is less disruptive than rotation of viewpoint and that, when subjects are blindfolded, actual rotation is less disruptive than imagined rotation. Again, this implicates the involvement and use of extra-retinal, proprioceptive or vestibular cues during actual movement of the observers. These cues could be used for instance to specify the relative direction and magnitude of rotation and translation and thus could support spatial updating in the absence of visual cues. Whilst the involvement of non-visual information in the visual perception of spatial layout is interesting in itself, it does not seem plausible that only such indirect information is used for spatial updating. Indeed any information which yields the magnitude and direction of the change in the viewers position could be used for spatial updating including indirect information derived from vision itself. Simons & Wang did test whether background cues were necessary for spatial updating by repeating their experiment in a darkened room with self-luminous objects. The results were only slightly affected by this manipulation. That is, spatial updating still occurred. However, this only means that visually derived movement is not a necessity for spatial updating. It does not imply that visually derived movement cannot facilitate it. To determine whether spatial updating can occur through purely visual sources of information we constructed a simple vision-only based spatial updating experiment using a virtual reality simulation to eliminate unwanted cues and depict the implied movement within a realistic setting. We attempted to replicate the conditions of the original Simons & Wang [12] experiment but

with simulated movement rather than real movement and within a simulated environment rather than a real environment.

1.1 The Principles of Virtual Environment Simulation

An alternative to using real-world scenes for studying spatial cognition is to use a virtual environment (or virtual reality) simulation. These are becoming increasingly popular means of simulating three-dimensional (3D) space for studying spatial cognition (e.g., [1,7,4]. Such simulation allows one to actively move through a simulated space with almost immediate visual feedback and also allows objects in the scene to move (either smoothly or abruptly) from one position to another.

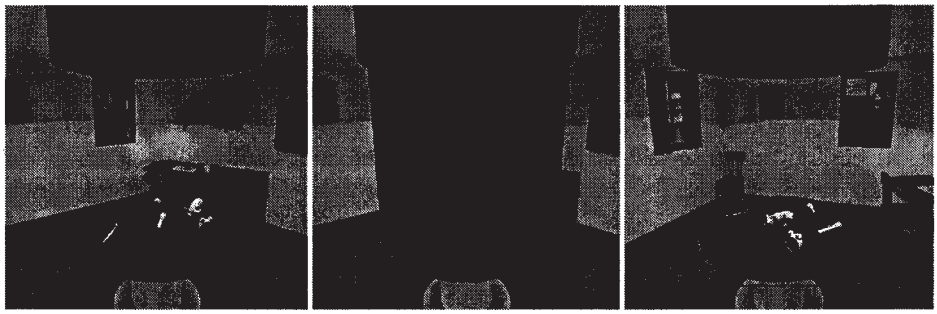

Fig. 1. Three views of the simulated environment marking the three stages of the experiment as seen by the viewpoint-change observers (see below). The first image depicts the 5 test objects. The middle image shows the fully lowered curtain. The right hand image shows the new view of the objects after a counter-clockwise rotation in view. In this case the 5 objects and table did not rotate so that the retinal projection is different from that in the initial view. The pig and the torch have exchanged places and the appropriate response would have been to press the 'change' button.

In essence, our simulation consisted of a 3D-modelled polygonal environment in which a virtual camera (the view of the observer) is translated and rotated and whose projection is rendered on a desktop computer monitor in real-time (see Figure 1). That is, the scene was rendered (i.e. projected onto the image plane and drawn on the monitor) approximately 30 times a second. A Silicon Graphics Octane computer performed the necessary calculations. The cameras motion was controlled by the observer using a 6 degrees-of-freedom motion input device (Spacemouse). The initial stages of development involved the construction of the 3D environment, created using graphics modelling software (3DStudio Max from Kinetix, USA.) The illumination in most VR simulations is usually calculated according to a point illumination source located at infinity. The visual effects produced by such a model are unrealistic because a non-extended light source at infinity produces very abrupt changes in illumination which can be confused

with changes in geometry or surface reflectance (lightness). Furthermore, in any real scene the light reaching the eye is a function of both direct illumination from the light source and indirect illumination from other surfaces. The latter helps to illuminate surfaces occluded from the light source and produces smooth gradients of illumination across surfaces which can also eliminate the confusion between reflectance and illumination mentioned above (see [3]). Therefore, in our experiments we pre-rendered the surfaces of our virtual environment using software that simulates the interreflective nature of diffuse illumination. This produced realistic smooth shadows and ensured that regions not visible to the source directly could still be illuminated by indirect light.

Once the 3D model was constructed, interactive simulation software could be used to control the simulated observer movement and subsequent rendering of visible field of view onto the screen. On SGI computers this is performed using the IRIS Performer 'C' programming library. This software also provides the functionality for detecting the simulated observers' collision with surfaces in the scene and makes sure the observer stays within the confined bounds of the simulation.

2 Method

2.1 Materials

The experiment utilised the 3D simulated environment described above together with 25 three-dimensional models of common objects. The objects were chosen for their ease of identification and consisted of, for instance, pig, torch, clock, lightbulb and toothbrush (see Figure 1). All objects were scaled to the same or similar size and no surface texturing or colouring was used. The objects could occupy any one of 7 possible evenly spaced positions across the platform and these positions where computed once at the beginning of each experiment. The position of each object for a given trial was chosen at random from the list of all possible locations. The object positioning ensured sufficient distance between each object so that the objects did not overlap and minimized the chances of one object occluding another from any given viewpoint.

2.2 Subjects

Male and female subjects were chosen at random from a subject database. All 28 were naïve as to the purposes of the experiment and had not performed the experiment in the past. They were randomly assigned to one of the two groups of 14 and given written instructions appropriate to the group chosen (see below). They were given an initial demonstration of their task.

Subjects viewed the scene through a viewing chamber that restricted their view to the central portion of the computer monitor and maintained a constant viewing distance of 80 cm. Before each block of trials they were allowed to move themselves freely through the environment for 3 minutes (using the Spacemouse)

to get a better impression of its dimensions and to acquaint themselves with the five test objects, which were visible on the platform. Observers' simulated height above the floor was held constant at an appropriate level. The experiment was completely self-initiated and subjects received on-screen instructions at each stage.

2.3 Procedure

In keeping with the original experimental paradigm of Simons & Wang [12] we facilitated a retention interval by the simulation of a cylindrical curtain which could be lowered to completely obscure the objects from view (see Figure 1). Each trial consisted of the following format. The configuration of objects would be viewed for 3 seconds from the start viewpoint. The curtain was then lowered, eventually obscuring the objects completely. The observers' view was then rotated to a new viewpoint (Group B) or rotated half-way to this new viewpoint and then rotated back to the start viewpoint (group A). This retention interval (during which time the objects were not visible) lasted for 7 seconds for both groups. The curtain was then raised revealing the objects for a further 3 seconds. Subjects then had to decide if one of the objects was displaced (to a vacant position) during the intervening period. They responded by pressing one of two pre-designated keys of the computer keyboard. The subject's response together with their response latency (calculated from the second presentation of objects) was stored for later analysis.

The experiment consisted of 5 blocks of 12 trials each. For each block 5 new objects were chosen and a new start position around the platform randomly selected as the start viewpoint. In 50% of the trials there was a displacement of one of the objects. Also in 50% of the trials, the platform was rotated by 57 degrees in the same direction as the observer. Thus, for group B (different observation point), when the platform was rotated and no object was displaced, exactly the same retinal configuration of objects on the table was observed (apart from the background). In the case of group A (same observation point), only when the platform was not rotated and no object displaced was the retinal configuration of objects exactly the same. Thus both groups had to determine displacement of one of the objects when the retinal projection of objects was the same or a fixed rotation away from the observation point. For group B, however, the rotation was the result of the observers simulated movement whereas for group A it was the result of the rotation of the platform and objects. The rotation angle of the simulated observation point and platform was always 57 degrees. For each block of trials the computer chose either a clockwise or anti-clockwise rotation away from the start observation point. This ensured that subjects paid attention to the rotation itself. When the table was rotated both groups were notified by on-screen instructions and received a short audible tone from the computer

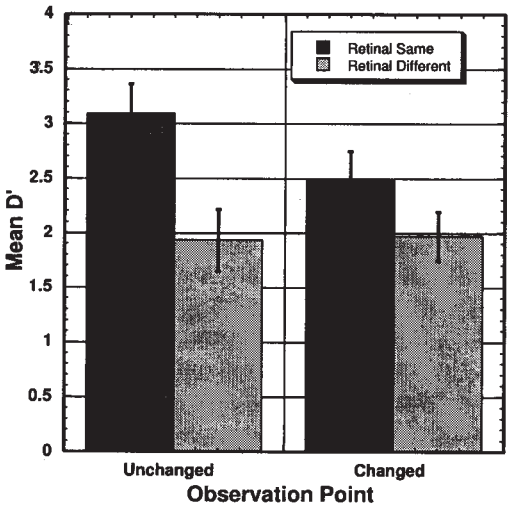

Fig. 2. Shows the mean d' for according to the two subject groups (which differed in terms of changed or unchanged observation) and same or different retinal projection. The error bars are standard errors of the mean.

3 Results

The proportion of correct responses and false alarms were used to calculate a mean d' score for each observer. On the whole mean d' was always above 1.75 which indicates that subjects found the task quite easy. An analysis of variance (ANOVA) with two between-subject factors (Group A or B, i.e. same or different observation point) and two within-subject factors (same or different retinal projection) revealed that the effect of group on d' was not significant [$F(1,26)=.76$, $p=0.39$], that the effect of the within-subject factor retinal projection was significant [$F(1,26)=22.76$, $p<0.00005$] and that the interaction between these two was approaching significance [$F(1,26)=3.56$, $p<0.07$]. These results are plotted in Figure 2 which shows that group B (changed observation point) were least affected by a different retinal projection of the objects. This is more strongly indicated by the response times (RT). A similar ANOVA on RT revealed again that the effect of group was not significant [$F(1,26)=0.65$, $p=0.4$], that the effect of retinal projection was significant [$F(1,26)=21.67$, $p<0.0005$] and also the interaction between these was also significant [$F(1,26)=19.3$, $p<0.0005$]. This interaction is portrayed in Figure 3 which shows that there was no significant difference in response times for group B for identifying configurations viewed from either the original or rotated observation point. This is in sharp contrast to the group A observers who required on average 400 ms more in order to perform this judgement after rotation of the objects.

These results for RT indicates apparently view-independent performance for Group B and view-dependent performance for Group A. However, these response

times were averaged for all responses regardless of whether the response was correct or incorrect. It may be that these differences correlate only with the different error rates for each of the individual conditions rather than reflecting differences in mental processing. We therefore extracted the response times corresponding only to correct responses and performed t-tests for related pairs of within-subject data and for each group. The mean RTs for group A were 1447 ms and 2090 ms for same and different retinal projections respectively. For group B, mean RTs were 1544 ms and 1654 ms for same and different retinal projections respectively. We found that these RT measures for group B were not significantly different [$t(13)=-1.08$; $p<0.29$] whereas the means for group A were significantly different [$t(13)=-5.14$; $p<0.0005$]. Therefore, the correct responses for group A reflect a possible difference in the amount of time required to perform the task correctly when the retinal projection was the same or different. For group B, who performed the task from a different observation point, there was no significant cost in response time regardless of whether the same or different retinal projection of objects was viewed.

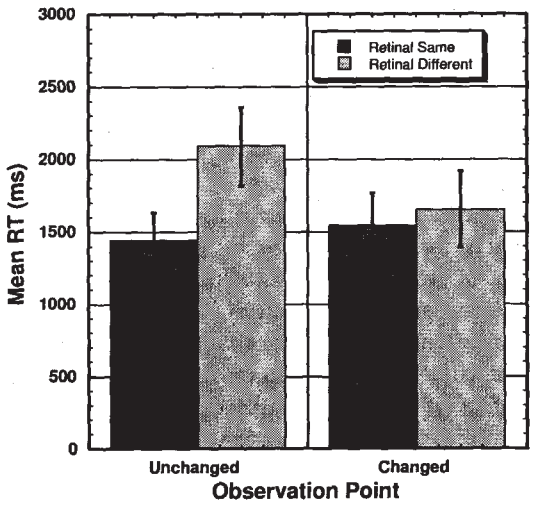

Fig. 3. Shows the mean response times (RT) for all responses (correct and incorrect). Error bars are again standard errors of the mean.

4 Conclusion

The ability to remember the relative spatial locations of several objects in a three-dimensional display has been used to reveal some of the properties of human spatial representation and the perception of spatial layout. One important

ability that requires spatial representation (and processing of spatial representations) is the ability to make judgements about relative locations of objects in a display after changes in the observers' viewpoint. Previous experiments suggest that such an ability is view-dependent and becomes view independent when additional information regarding the magnitude of the orientation shift is available from proprioceptive or vestibular sources [12,14]. We wanted to determine if this additional information could be derived solely from visual sources.

Informal responses from our subjects in the current experiment suggest that the principle means of performing the task was to remember a verbalised sequence (such as, pig, torch, clock, lightbulb and toothbrush). However, this method of storage still implies some directional encoding of the next object in each sequence. That is, it must encode the relative direction of the next object in the sequence with respect to the previous object. Furthermore, we may also suggest that the position of the start object in the sequence (i.e. pig) needed to be stored (perhaps in a world-centred reference frame) as well. Otherwise, if the start object was indeed the object that moved then this sequence encoding would be difficult to use. All this suggests that one or more reference frames were in used to encode the positions of objects and that differences either in the mode of encoding or operations on these encodings were the cause of the differences between groups that we have observed.

For conditions where objects rotated and the observers' observation point remained constant we found that performance dropped dramatically compared with trials in which the objects did not rotate and the observation point was also constant (group A). This misalignment of objects resulted in more errors and longer response latencies. This is in keeping with view-dependent results obtained in both object recognition (e.g., [2,13]) and scene recognition studies [5,4]). If either a retinal-centred or body-centred encoding was used these orientation changes required some computational effort or transformation to extract the necessary matching information. One means of performing the task is to imagine looking at the objects from a new position which preserves the original retinal-centred encoding. In this situation, one could determine the movement of an object by comparing two similar representations. If this is the case however, then the results of previous studies on imagined spatial layout would have predicted the loss in performance (e.g., [8]) which we obtained.

For the subjects in the view-change group (group B) the differences between same retinal image and different retinal image were found to be statistically insignificant. Again one strategy for subjects would have been to imagine they were at the original viewpoint before the move and compare relative spatial positions in a retinal-centred reference frame. However, what they may also do here is continually update the possible changes that are occurring to the visual appearance of objects as their viewpoint changes. This may explain why in the original experiments by Simons & Wang viewpoint-change subjects performed better with different retinal projections of objects than with the same retinal projections (we would expect that matching two dissimilar retinal images results in more errors, not less). The operation of this predictive mechanism may be

initiated automatically whenever we move around the world. The fact that there was no real movement in our present study may explain why we did not obtain similar elevated performance by Group B subjects. In future experiments, this could be tested by comparing abrupt and smooth changes in view and also by observing the influence of actual walking but with visual feedback derived from a head mounted display.

As mentioned above, our computer simulation differed from the previous experiments in one crucial factor. Namely, that in this simulation observers' movement was implied rather than actual (observers passively viewed their rotation around the table). In our experiment therefore, the key information supporting the task was entirely visual. At first it appears that this is contradictory to the results of Experiment 2 reported by Simons & Wang [12] in which they performed the same task after removing the visual detail in the background. The background detail was reduced by performing the experiment in the dark and with self-luminous objects. Simons & Wang found that this did not affect spatial updating and the superior performance of view-change subjects. However, the fact that the absence of visual background does not affect performance does not imply that visual background cannot provide the necessary information. Furthermore, in our experiment the superior performance of viewpoint-change subjects is reflected principally by differences in their response times, which were not recorded in Simons & Wang's experiments. It appears that the crucial factor is the provision of additional information regarding the motion or change of position of the observer. This additional information may be derived from various sources, both visual and non-visual.

In conclusion, this experiment has revealed a positive benefit of simulated movement even within a simulation of a three-dimensional space and even when only visual input specifies that the observer has moved. This indicates that information from both visual and non-visual modalities can interact in the facilitation of stable perception and view-invariant recognition. The basis of this facilitation appears to be that of providing a stable environment or 3D space and a continual mapping of the position of the observer within this space. Future studies may determine whether this does indeed imply spatial updating of the observers' relative position in space and perhaps suggest some model of how spatial updating functions in the perception and recognition of spatial layout.

References

1. V. Aginsky, C. Harris, R. Rensink, and J Beusmans. Two strategies for learning a route in a driving simulator. *Journal of Environmental Psychology*, 17:317 – 331, 1999.
2. H. H. Bülthoff and S. Edelman. Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60 – 64, 1992.
3. C. Christou and A. Parker. *Simulated and Virtual Realities: Elements of Perception*, chapter Visual realism and virtual reality. Taylor & Francis, London, 1995.
4. C. G. Christou and H. H. Bülthoff. View dependency in scene recognition after active learning. *Memory & Cognition*, 27:996 – 1007, 1999.

5. V. A. Diwadkar and T. McNamara. View dependence in scene recognition. *Psychological Science*, 8(4):302 – 307, 1997.
6. M. J. Farrell and I. H. Robertson. Mental rotation and the automatic updating of body-centered spatial relationships. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(1):227 – 233, 1998.
7. E. A. Maguire, N. Burgess, J. G. Donnett, R. S. J. Frackowiak, D. D. Frith, and J. O'Keefe. Knowing where and getting there: a human navigation network. *Science*, 280(1):921–924, 1998.
8. C. C. Presson and D. R. Montello. Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception*, 23(12):1447–1455, 1994.
9. J. J. Rieser. Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15(6):1157–1165, 1989.
10. J. J. Rieser, A. E. Garing, and M. F. Young. Imagery, action, and young childrens spatial orientation - its not being there that counts, its what one has in mind. *Child Development*, 65(5):1262 – 1278, 1994.
11. A. L. Shelton and T. P. McNamara. Multiple views of spatial memory. *Psychonomics Bulletin & Review*, 4(1):102 – 106, 1997.
12. D. J. Simons and R. F. Wang. Perceiving real-world viewpoint changes. *Psychological Science*, 9(4):315 – 320, 1998.
13. M. Tarr and H. H. Bülthoff. *Object recognition in man, monkey and machine*. MIT Press, 1999.
14. R. X. F. Wang and D. J. Simons. Active and passive scene recognition across views. *Cognition*, 70(2):191 – 210, 1999.