Max–Planck–Institut
für biologische Kybernetik
Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 53

View-direction specificity in Scene Recognition after Active and Passive Learning

Chris Christou[1] & Heinrich H. Bülthoff[2]

October 1997

Chris Christou is supported by a scholarship from the Max-Planck Gesellchaft. Many thanks to Jeff Liter and Fiona Newell for comments and suggestions during the preparation of this document.

[1] Max-Planck-Institute fuer biologisch kybernetik, E–mail: chris.christou@tuebingen.mpg.de; to whom all correspondence should be addressed
[2] Max-Planck-Institute fuer biologisch kybernetik, E–mail: heinrich.buelthoff@tuebingen.mpg.de

This document is available as /pub/mpi-memos/TR-53.ps via anonymous ftp from ftp.mpik-tueb.mpg.de or from the World Wide Web, http://www.mpik-tueb.mpg.de/bu.html.

View-direction specificity in Scene Recognition after Active and Passive Learning

Christou C.G. & Bülthoff H.H.

Abstract. Human visual encoding of scenes was studied with respect to view specificity: the extent to which observers ability to recognize a familiar scene depends on the limited set of directions experienced during training. Precise control of cues was achieved by using a computer model of a virtual room. A novel explorative search paradigm was implemented using real-time image generation and provided a controlled yet natural means of familiarisation.
After training, observers where able to recognize both familiar and novel direction views but the latter involved more errors and required more processing time. Observers were also able to identify the corresponding floorplan indicating that the encoding could provide abstract relation information.
In subsequent tests in which we replaced the interactive training stage with passively observed sequences of snap-shots we found that although familiar view recognition persisted the novel view generalisation performance diminished. This was so, even when binocular disparities were used to provide more depth information.
These results suggest that mental encoding of scenes is view-based although the observed generalisation to novel views cannot readily be explained by simple transformations on 2D stored views. Furthermore, our study highlights the importance of allowing natural behaviour during visual familiarisation tasks.

---

1 Introduction

The fact that we have the ability to remember scenes and make predictive navigational decisions within them indicates that we have the use of some form of visually encoded representation at our disposal. Current theories regarding the form taken by this visual encoding include structural descriptions based on generic parts (e.g. Marr & Nishihara; Biederman, 1987) and sets of multiple 2D views (Tarr and Pinker, 1989). These theories make different predictions regarding behavioural responses to be expected from an observer using a particular form of encoding for recognition (see Pinker 1988 for an introduction). The structural description theories, for example, state that recognition ability is unaffected by the particular view one has of a familiar object; that is, the representation is view independent provided that sufficient structural detail is visible. Results from recognition studies using familiar objects support this prediction (Biederman & Gerhardstein, 1993). The multiple-views theories on the other hand would predict that when a particular stimulus does not match a 'stored-view' of an object then recognition performance (measured in terms of error rates and response latencies) will depend on the deviation of the current view from stored views Tarr &Pinker 1988; Bülthoff and Edelman, 1992) Computational models have been developed which show how a strictly 2D view-based representation can in principle be used to recognize 3D objects under various transformations (e.g. Poggio & Edelman, 1990; Ullman & Basri, 1991).

In this paper we report on experiments which investigate view-specificity in the recognition of 3D scenes. Most view dependency experiments have involved 3D isolated objects because visual cues are easy to control and, as is the case more recently, using computer graphics one can generate any number of novel structured stimuli one wishes. The few scene recognition studies that do exist have, in contrast, been carried out in real environments. The findings of these studies are not clear cut regarding view-dependency and generalisation to novel views. Shelton and McNamara (1997) have recently presented evidence for view-point specificity in the perception of spatial layout. They exposed observers to the spatial configuration of objects in a room but from only two viewpoints. Subsequent testing on observers ability to judge the relative spatial location of each object found errors were lower when relative heading judgements were made from the two familiar directions. On the other hand, Hock and

Schmelzkopf (1980) presented what they believe is evidence for view-independent representation. In their study they took multiple photographs of a scene from various vantage points and, after training, observers had to associate these photographs with a single vantage point. Observers also had to associate unfamiliar photographs with these vantage points. They found that observers were able to do this grouping and also to pin-point the vantage points on schematic diagrams of the scene. They took this as evidence that observers had acquired view-independant schematic knowledge of the scene. Similarly, Rowland,Franken,Bouchard, & Sookochoff (1978) found that viewing a scene from an increasing number of directions improved generalisation to novel views. The ability to recognize individual exemplar images of the scene was, however, reduced after long intervals. This suggests that the representation, which may be view-specific initially, becomes more generalised and schematic over time.

We believe that there are two main problems with the previous studies of scene recognition that may have unfairly biased results either towards view-dependency or an unwarranted degree of generalisation to novel views. First, under natural conditions spatial encoding occurs during locomotion. Studies which limit viewing to just one or two viewpoints in a scene (such as Shelton and McNamara, 1997) are unfairly biasing observers toward view-dependency. On the other hand studies without some control of observer movement would be able to conclude very little regarding generalisation to novel views. Clearly some form of natural yet precisely restricted movement within a scene is desirable as it preserves the natural means by which scenes are encoded. Secondly, the majority of the scene recognition experiments conducted to date have used real world scenes or photographs of real world scenes in training. This produces problems of cue control. For instance, the observers in the experiment of Hock and Schmelzkopf (1980) may have used the similarities in illumination depicted by each photograph as a grouping principle (assuming photographs were taking at the same time of day.) In general, subsidiary cues such as landmarks or features such as colour or textures may serve as indexing cues, which could, falsely, suggest a generalised representation. In order to address such problems we used computer graphics for environment modelling together with real-time image generation to simulate natural motion through a virtual scene. The experimental question was whether after naturalistic familiarisation with an environment from a limited set of directions observers recognize the environment when viewed from novel perspectives. If subjects store a structural, view-independent, representation we would expect to see little differences when comparing novel view with familiar view performance as measured in terms of response latencies and error rates. If, on the other hand, some transformation of the internal representation takes place, then response latencies should be higher for novel views but error rates should be as low as for familiar views (see Bülthoff, Edelman and Tarr, 1994). If no explicit representation of the novel direction views has been formed then generalisation to novel perspectives should be difficult, if not impossible, even after natural encoding has taken place.

1.1 The Setting: A virtual attic in a virtual house

The test environment was a computer model of the attic of a real house situated in Tübingen, Germany (see figure 1). The irregular layout of this house resulted in visually rich structural detail and made an interesting setting for an experiment requiring restricted use of furniture and textural detail (the latter could be highly informative for recognition purposes.) The entire house was modelled using the Medit modelling system on a Silicon Graphics Indigo II High Impact computer and consisted of approximately 6000 polygonal elements. We used architectural drawings as a basis and onsite visits provided further details. However, the specific requirements of the experiments meant that some novel detail had to be introduced.

As a distractor environment the polygonal constituents of the computer graphics model of the attic were mirror reflected about a vertical plane (along the main corridor) and structural alterations where made to the resulting model so that the same large-scale features were preserved but with distinctive modifications (see figure 1c).

Both target and distractor models had the same 4 relative light sources assigned to them using SGI Performer graphics library commands. Surface shading was simulated using a constant-elements lighting model in which the intensity is calculated once for the entire extent of each polygon based on its orientation with respect to the light sources. The lighting model did not use occlusion tests, so no cast shadows were simulated. To increase structural discrimination, different colours were used for the floor, walls, simulated wooden beams and bannisters.

We also used the Performer API library to sim-

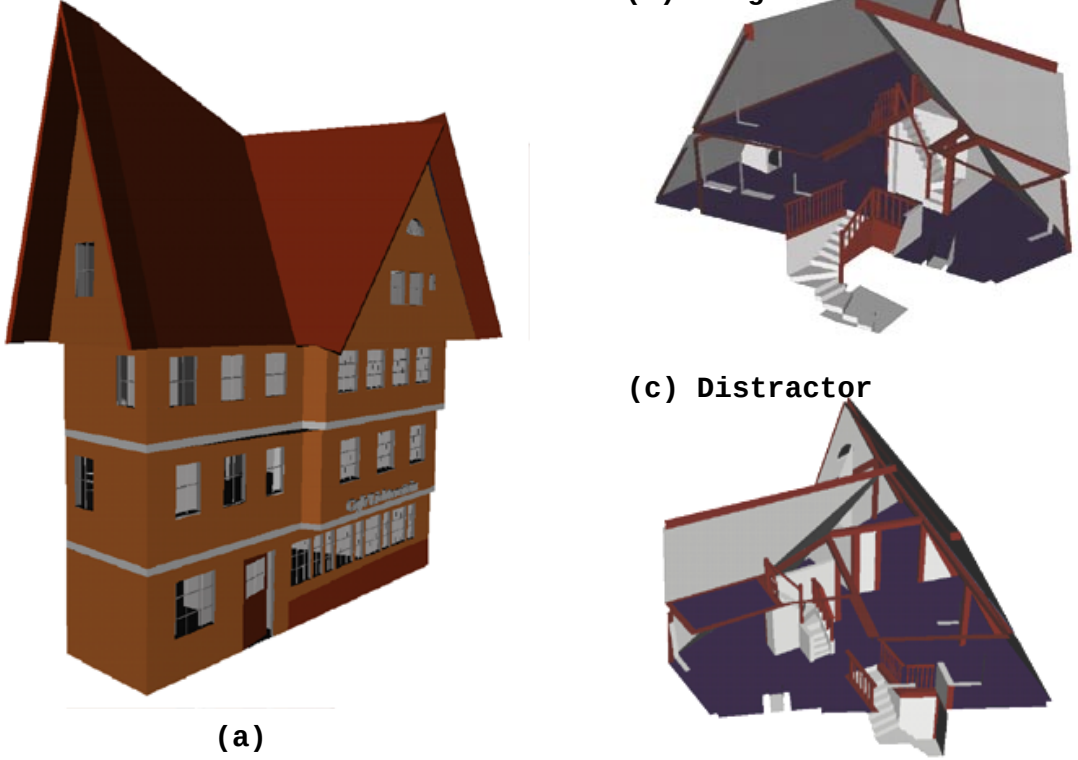

Figure 1: The computer model of the Tübingen house (a) together with an unimpeded view of the attic (b) which provided the setting for these experiments. The distractor (c) was derived from the target, but structural changes were made to make the two distinguishable.

ulate movements within the model. These movements (which included translation and changes in heading and pitch) were made by the observers via manipulation of a SpaceBall (Spacetec IMC Co., Massachusetts, USA). The SpaceBall is a simultaneous six-degree of freedom desktop motion control device, which consists of a latex pressure sensitive ball attached to a hand rest. The image of the scene changed in real-time according to the direction of force or torque applied to the ball. Movement restrictions were enforced when appropriate and auditory feedback (in the form of a high frequency tone) was given when necessary (see movement restrictions below).

2 Experiment 1 - Generalisation to Novel Views After Active Learning

2.1 Subjects

In total, 9 men and 9 women between 19 and 35 years of age participated in the experiment in return for pay. All observers were naive as to the purpose of the experiment and had never seen the computer model or real attic interior used for learning. All experiments were completed within 70 minutes (including training). All observers were given the same printed instructions before training and tests.

2.2 Procedure

2.2.1 Learning Phase

Observers familiarised themselves with the computer model whilst seated 80 cm in front of a 75Hz RGB monitor (whose outline was occluded) in a darkened room with their preferred hand placed on the SpaceBall. Their movements through the model were presented in a 45° field of view window consisting of 600x600 pixels.

To motivate them to learn the model in a natural manner, we asked them to search for several two-digit codes (markers) placed in various 3D locations within the model. During their search for markers, the observers simulated movement was restricted. The particular viewing restrictions used are illustrated in figure 2. Firstly, subjects were only allowed to manoeuvre within a rectangular region as illustrated in the topographic view of figure 2a. This resulted in a smaller, and therefore easier, search domain. Secondly, the observers simulated height was restricted to 1.75 m. Thirdly, heading and pitch changes were restricted to 60° about a free vector along the length of the translation corridor. This effectively meant that observers could not look back during their search. They were however allowed to press a reset but-

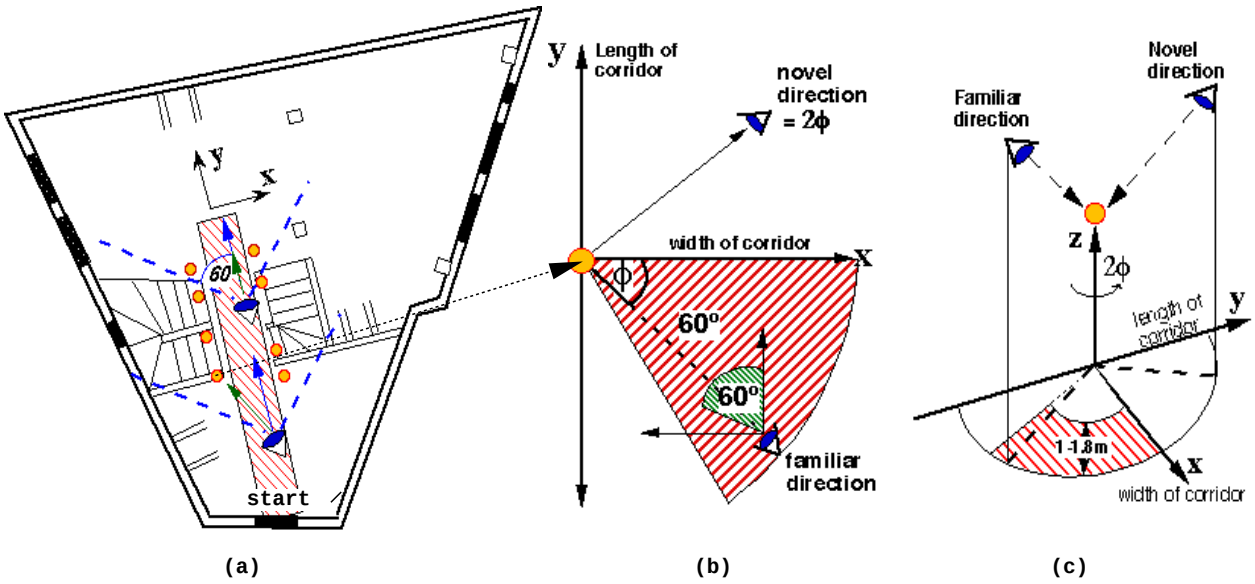

Figure 2: (a) Floorplan of the training environment. The observers' simulated movements were restricted to the hatched rectangular area. The coded markers (filled circles) were positioned on either side of this area. Observers initially looked along the Y axis and could change their simulated heading but only by 60° to the left or right. This ensured they could never look back. (b) Illustration of these heading restrictions. The markers became visible only when observers were within an appropriate distance (1.0 - 1.8m) and subtended an angle ϕ less than 60°. (c) Shows how this angle ϕ was used to generate the novel direction views. The novel viewing point was obtained by rotating about the vertical (Z) axis in the horizontal plane by 2ϕ and appropriate changes made to the heading such that the marker position was now observed from a novel direction.

ton on the SpaceBall at any time they wished to return to a fixed starting point (see figure 2a).

The markers that observers searched for had constant size independent of simulated viewing position and could in principle be occluded by other objects. Initially the markers were invisible until the observer came closer than a threshold distance at which point the markers appeared as two colour-coded stars (red = too close, blue = too far away, green = good position but inappropriate viewing angle.) The actual two-digit codes would become visible only when the observer had an appropriate viewing distance (approx. 1m) and satisfied an obliqueness of view requirement. These requirements ensured that a variety of different, 'close-up', views formed the stimulus set. The obliqueness of view requirement was implemented by ensuring that the angle ϕ between the vantage point and the X axis (see figure figure 2b) was less than 60°. Observers were instructed to manoeuvre themselves such that the two-digit codes were in the centre of the screen before entering the number into the computer via a number pad located on the SpaceBall. At this point their virtual viewing position in 3D together with heading and pitch angles were stored together with the marker location for stimulus generation. There were 14 such targets in total and all observers were required to find all of them. The explorative learn phase was completed by all subjects within 20 to 25 minutes.

Because none of our observers were familiar with the SpaceBall, all observers spent an initial 10 minutes performing the above task in an unrelated training environment. Only after successfully finding 7 markers in this environment could they proceed to the learning phase.

2.3 Test Phase

After completing the learning phase observers were instructed to look away from the screen during which time 2D images of familiar and novel views were generated from their stored data (see examples in figure 3). The familiar view was simply an image of the location of each marker as defined by the observers vantage point, heading and pitch. The corresponding novel direction view was generated by rotating the vantage point by 2ϕ (clockwise or anticlockwise) about the vertical axis

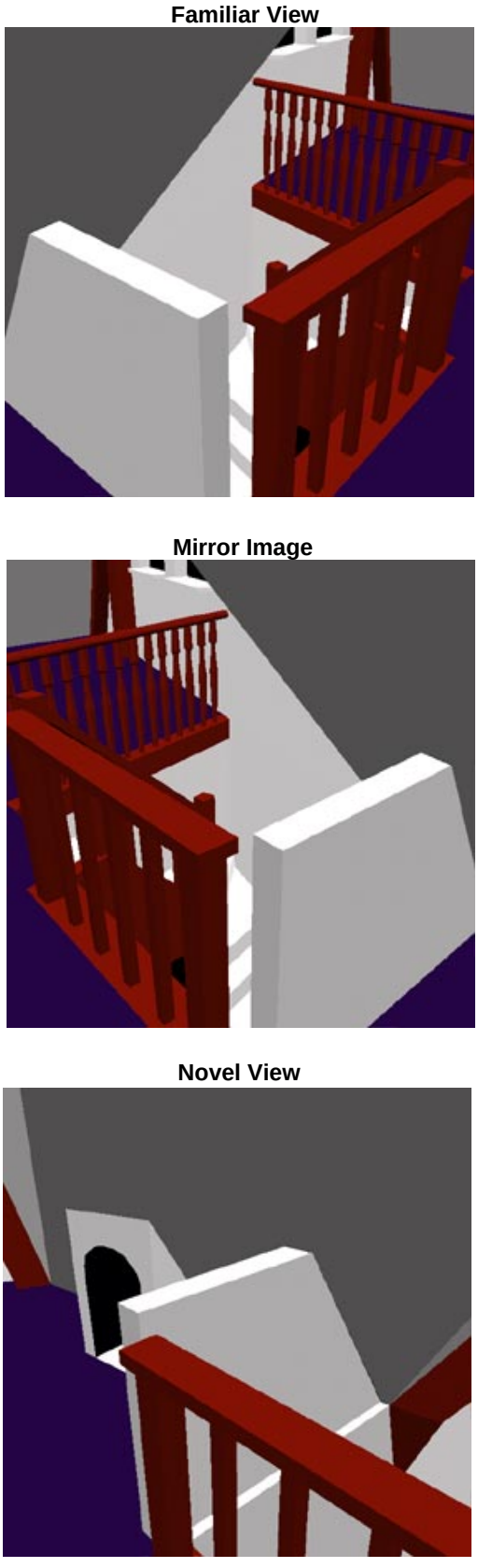

Figure 3: Example images used as stimuli in Experiment 1.

through each marker where ϕ is the angle defined above (see figure 2c). Thus, the rotation angle varied according to each observers settings but was always greater than 60°.

As a simple test for recognition by feature analysis we also included the horizontal mirror reflected image of each familiar-view stimulus (that is, swapping corresponding pixels on either side of the vertical medial bisector of the image). If a simple feature analysis is sufficient for recognition purposes then these images should be recognized significantly often. Finally, each of the three stimuli were matched with a corresponding image taken from the 3D distractor environment described previously.

Before performing the recognition tests, each observer was required to identify the topographic floorplan that showed the most correspondence to the 3D learned model. Topographic map usage has often been studied experimentally in relation to human navigation (e.g. May, Peruch and Savoyant, 1995) although here we use it as a means of testing the ability to make abstract judgements of layout immediately after training.

In all there were 8 choices of map printed in colour on a white sheet of A4 paper (see figure 4). Only one of these corresponded closely to the model in terms of gross detail. All floorplans were line-drawing derivatives of the attic or the distractor model. For both attic and distractor we generated mirror reversed floorplans by simple line element mirroring about a fixed coordinate frame. Slight structural modifications were then made to the target and distractor geometries to produce 6 other 'distractor' floorplans. All plans were then randomly rotated in the image plane. These operations produced 8 variations, only one of which matched the target attic. In all cases coloured arrows indicated the up/down direction of the stairs.

Observers performed the topography test two times; once, immediately after the learning phase and then again after the recognition tests. On the occasion of the second test they also had to identify the correct floorplan for the distractor that was seen in the picture recognition tests. This was to check whether observers also formed an abstract model of the distractor during testing.

After the floorplan test observers were presented with the 2D picture stimuli taken either from the target or the distractor model. The dimensions of the stimuli were the same as in the training phase (example stimuli taken from the target scene are shown in figure 3. A yes-no design

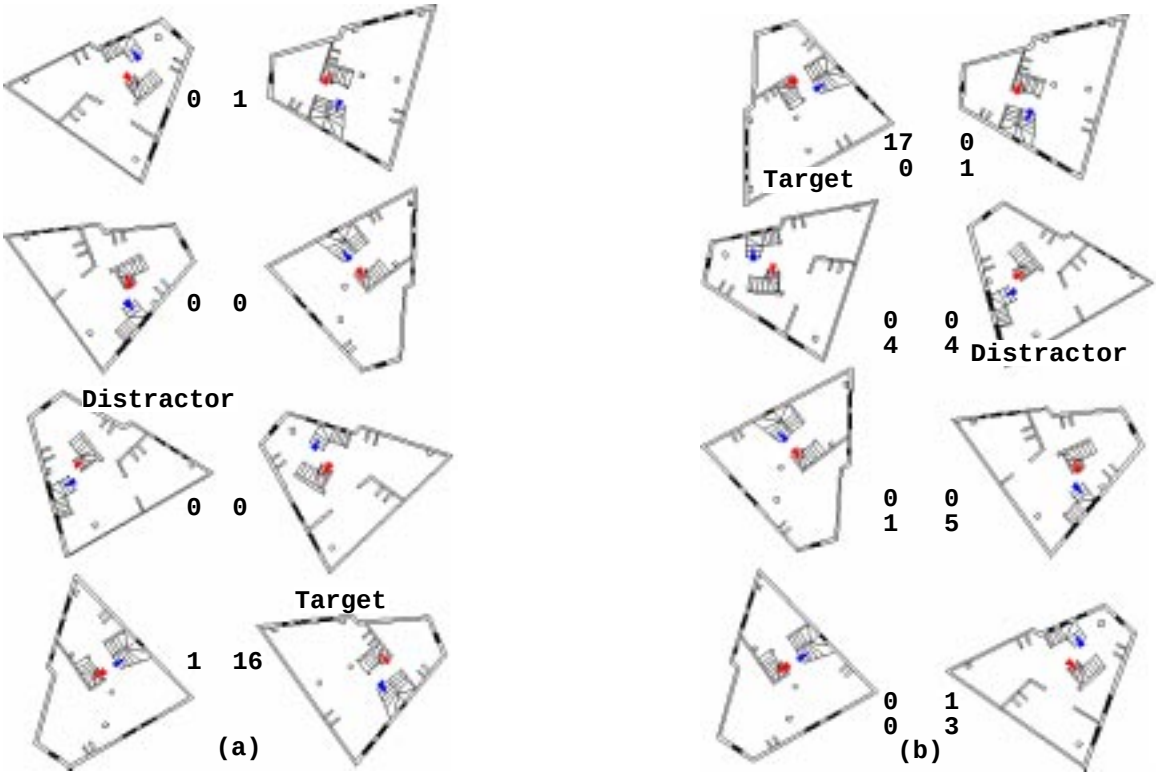

Figure 4: Number of times observers selected each floorplan as corresponding to the training environment. (a) First test - directly after training. (b) Second test - after picture recognition tests (Numbers below signify identification of the distractor environment).

was used in which a single stimulus remained on the screen until the observer responded yes or no using input buttons on the SpaceBall. Observers were instructed to respond yes if they believed the picture showed a view of the learned model and no otherwise. They were informed that the pictures could be taken from any direction within the scene. Observers were further instructed that they should be 'as fast and as accurate as possible.' Response latencies were measured from stimulus onset to time of response.

2.4 Results

The proportion of times each floorplan was identified as that of the familiar model on the first test is shown in figure 4a. In total 16 out of 18 (or 89%) observers identified the topography of the training model. Furthermore, when asked, observers could also give an indication of their general movements in terms of the diagram. When asked how they came to their decisions, the predominant response was that some relational information was used such as the location of the windows with respect to other scene components such as fireplaces

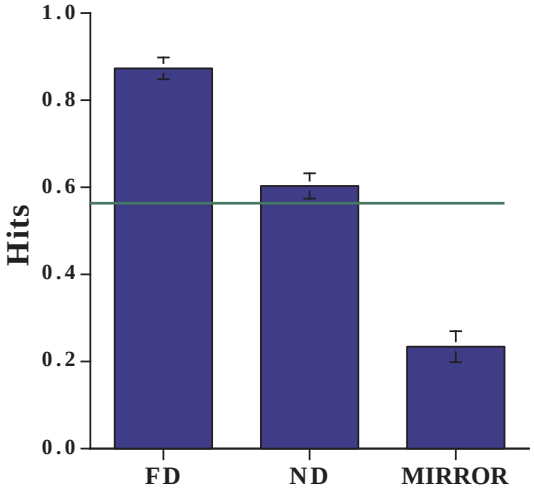

Figure 5: Proportion of hits for each image type: FD = images of marker locations taken from familiar directions; ND = images of marker locations taken from novel directions; MIRROR = mirror images of FD. The error bars correspond to the standard error of the mean for all 18 observers.

or stairs. After the picture recognition test observers again had to identify the correct floorplan as well as the 3D distractor environment used in the picture tests. The floorplans used, depicted in figure 4b, were the same as in figure 4a, but in different locations. Results were consistent with the first test with 17 out of 18 observers (or 94%) making the correct choice. Scores for the distractor on the other hand were almost equally distributed among the 4 closest floorplans.

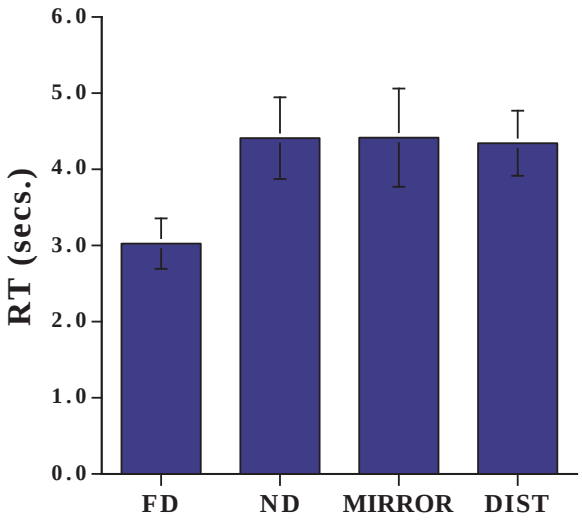

Figure 6: Mean correct response times for each image type in Exp. 1. The error bars correspond to the standard error of the mean.

The data for the picture recognition tests were analysed in terms of proportions of hits and false-alarms. Since all stimuli (including distractors) were presented in a single block, the false-alarm rate (.26) is calculated as the average for the three stimulus conditions. An analysis of variance on the hit rates with gender a between subjects factor and image derivation [i.e. Familiar direction view (FD), Novel direction view (ND), and Mirror image (MIRROR)] a within-subjects factor revealed an overall significant effect of image derivation ($F_{2,32} = 99.5; p < .0001$) but no effect of gender ($F_{1,16} = 0.15; p > .5$). There was also no clear interaction between these two ($F_{1,32} = 0.55; p > .25$)

The proportion of hits for each stimulus condition is shown in figure 5. The horizontal line marks the hit rate at which we can be 90% confident that the corresponding d' (given the false alarm rate) does not include the value of zero (MacMillan & Creelman, 1990). The mean proportion of hits for FD [mean=0.87, median=0.89, corresponding to mean $d' = 2.0$ (average over all observers)] and for ND (mean = .60, median= .61, $d' = 0.93$) are both above thus line. This is not the case for mirror images (mean=0.23, median=0.21, $d' = -0.21$). The minimum and maximum d' for condition ND was 0.2 and 1.8 respectively.

Response times for correct identification of FD images were also significantly faster than they were for the other two conditions and the average correct rejection time (see figure 6). An ANOVA showed an overall significant main effect of response time ($F_{2,32} = 5.02; p < .025$). A post-hoc analysis (Scheffe test) revealed that the FD views were recognized significantly faster than the ND views ($p < .05$) and also the mirror images ($p < .05$). However, there was no difference between ND views and mirror images ($p > .5$)

The results of this experiment therefore confirm a strong dependence of recognition on the familiar set of viewing directions observed during training; even though natural (yet restricted) movement was allowed. However, many observers could clearly recognize the novel direction views and this is rather surprising given that these views were never experienced. We can be sure that these views were never experienced because of the no-looking-back viewing restriction. The corresponding novel views of marker locations often revealed new structural detail or lacked previously seen detail. Thus, a simple transformation of stored FD images would not suffice for recognition. To illustrate this point, consider the familiar direction view and corresponding novel direction view in figure 3. The overall hit rate for the familiar direction was 92% (above average). The hit rate for the novel direction was 62% (average). This means that even though new detail was introduced (the fireplace) the image was still identified as a depiction of the target and not the distractor.

There is evidence that allowing observers natural movement in 3D scenes facilitates a better ability to compensate for any transformation in the visual field. For instance, Wang and Simons (1997) have found that observers compensate better for layout changes of objects when viewing changes are brought about by ego-motion than by rotation of the objects in the scene. We were therefore interested to see how much of this ability to generalise to novel views was facilitated by allowing observers to move within the scene. This has consequences for visual encoding because if visual encoding takes the form if stored-views then there should be no difference between an active (mobile) observer and one who is presented with a series of snap-shots during familiarisation.

3 Experiment 2: Recognition after Passive Learning

3.1 Subjects

Eighteen observers (9 men and 9 women, all naive) between the age of 19 and 35 years participated in the experiment in return for paid reward.

3.2 Procedure

3.2.1 Training Phase

Again, the procedure utilized a training phase and a test phase. The latter was identical to that used in Experiment One. In the training phase observers sat in front of a computer monitor while a series of 50 images were repeatedly presented to them. Each image was presented for 1.5 seconds with a 1 second interval between each image. Fourteen of these images were obtained from the settings of the best performing observer in Experiment 1 (in terms of d' and mean RT). The other 36 images were snap-shots of the model taken under the same movement restrictions applied previously. These images also consisted of wide-angle views taken from the starting point. Effectively all of the model that could have been seen by the previous observers was represented here.

In order to motivate observers to be vigilant and to emphasize the target views, each of the 14 target images contained an embedded two-digit code as in the previous experiment. Observers were instructed to press the reset button of the SpaceBall when they saw such a code. This stopped the sequence and they had to enter the code into the computer. Once the code was verified the sequence was resumed. The number of codes visible during any sequence of presentations was randomized between 1 and 4. The sequences were repeated until all codes were detected. This meant that there were always at least 4 sequences used for each training session. In actuality, this training phase lasted for between 20-25 minutes for each observer, equivalent to the training period of the previous experiment. The subjects were told to try to form a good impression of the model depicted because there would be a subsequent recognition test. The images were arranged in sequence such that no two images of the same region of the model were viewed in succession. This reduced the possibility of apparent movement through the scene.

3.2.2 Test Phase

The test procedure was exactly the same as in Experiment One. After training, observers were asked to identify the topography of the target model and then perform the picture recognition test. The test images were the same for all observers. The stimuli again consisted of the 14 familiar views of marker locations together with the corresponding novel views and mirror images. Finally, observers again had to identify both the target model floorplan as well as the distractor.

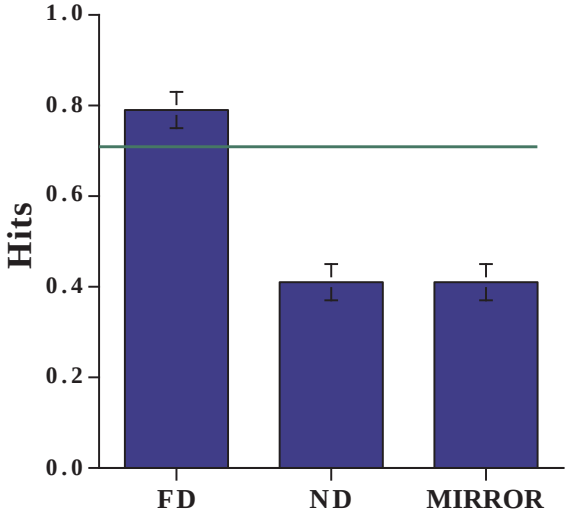

Figure 7: Proportion of hits as function of image type for Exp. 2.

3.3 Results

In total, the correct floorplan was identified 14 out of 18 times. This is a slight deficit compared to experiment one (16 identifications) although this still constitutes the majority of observers. On the second trial after the picture tests this fell to 13 out of 18 correct identifications. The distractor on the other hand was identified only two times, with the majority of choices spread between the three other most similar alternatives.

The mean hit-rate for each condition in the picture recognition tests is shown in figure 7. The mean false alarm rate was 0.34 ($s.e.$ = 0.028). The mean hit rate for FD was highest at 0.79 (median=0.79, $d' = 1.58$); for ND it was 0.41 (median = 0.43, $d' = 0.2$); for the MIRROR condition it was 0.41 (median = 0.40, $d' = 0.14$). For the novel direction views the minimum d' score was -0.6 and the maximum was 2.2. However, the observer responsible for this latter score was the only one to produce a $d' > 0.75$. An ANOVA with image type as a within-subjects factor revealed an overall significant difference between the respective hit rates ($F_{2,34} = 31.68; p < .0001$). Using a post-hoc analysis (Scheffe test) we found that

this difference was entirely due to the higher hit rate for the familiar views. The difference between the mirror images and the images of novel views were not significantly different. A between-subjects t -test on d' showed no significant difference between the FD condition results of Experiment One and those of the current experiment ($t_{34} = 1.29, p > .2$), whereas a similar test on ND showed that the d's for Experiment One were significantly higher($t_{34} = 3.75, p < .001$).

The mean response times for correct identifications follow a similar pattern to experiment one (see fig. 8). The difference in response time for the three conditions was significant ($F_{2,34} = 11.41; p < .0005$). A post-hoc analysis showed that the familiar view RTs were significantly faster than the RTs for both the novel direction views ($p < .0005$) and the mirror images ($p < .005$).

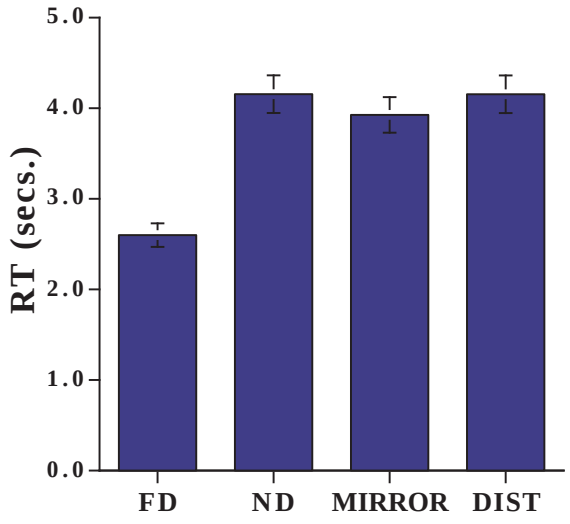

Figure 8: Mean correct response times for Exp. 2.

It appears that generalisation to novel views was more difficult to achieve in this experiment. If the generalisation to novel views observed in the previous experiment was achieved by transformations on (stored) 2D views (i.e. by transforming either the stored representation or the input stimulus) then this would not be expected. We must therefore ask what additional information is afforded to the mobile observer. The most obvious advantage is the increased availability of 3D cues that arise, for instance, from motion parallax and changes in interposition. These cues would be useful in reducing depth ambiguities, would improve scene segmentation and provide better distance information. In order to test whether lack of 3D detail contributes to the difficulty of this task the third experiment compared two conditions in which the only difference was the availability of additional depth information.

4 Experiment Three: Influence of Depth Cues

4.1 Methods

4.1.1 Stimuli

A powerful cue to 3D depth is derived from binocular stereopsis, which yields retinal disparities that scale with the relative distance of objects. In this experiment, retinal disparities were introduced by pairs of temporally overlapping disparate computer images whose presentation was synchronized to the opening and closing of the left and right shutters of a pair of CrystalEyes LCD shutter glasses. The effective framerate was 60 Hz. The left and right half-images were produced by moving through the computer model and taking stereo 'snap-shots' of different locations with a simulated interocular separation of 6cm (constant for all observers). The average viewing distance was 1.5 m. The resulting images contained both positive and negative (converging and diverging) disparities. The image dimensions were the same size as in the previous experiments.

4.1.2 Subjects

Observers where randomly assigned to one of two groups; MONOCULAR or BINOCULAR. Each group contained 16 observers matched for gender. One group was trained using binocularly presented disparate images of 16 target views and 34 additional views of the computer model. The second group was trained with the same series of images but using only their preferred eye (the other was masked by an eye-patch). Other details were the same as in previous experiments.

All observers regardless of group were given a simple stereo test (Stereo Optical Co., USA) and had to achieve a stereo acuity better than 40 arcsec. in order to participate. Two observers (both originally assigned to the disparate group) failed this test and were reassigned to the non-disparate group.

4.2 Procedure

4.2.1 Learn Phase

The training followed the same routine as in the previous experiment. Both groups of observers wore the stereo shutter glasses regardless of experimental condition.

4.2.2 Test Phase

The test phase followed immediately after training. Observers first identified the topographic map of the target model then performed the picture recognition task. The floorplan identifications were limited to 4 minutes, after which time observers were told to guess.

The images used in the picture tests were the monocularly presented left or right half-images derived from the training set, which contained a coded marker, together with novel views of the corresponding fixation points in 3D. To make the comparison between the two groups more fair the stimulus presentation time was limited to 1.5 seconds for both groups.

4.3 Results

The performance for identifying the topographic maps for both conditions was severely disrupted by the time limitation on responses. In all, only 6 observers out of 16 recognized the model after monocular training. This improved to 11 out of 16 after the picture recognition task. For the binocularly trained observers, 10 out of 16 recognized the model on their first attempt and 11 out 16 on their second. The apparent improvement for monocularly trained observers reveals a reinforcement after seeing the model again during the picture tests.

The overall false alarm rates for BINOCULAR and MONOCULAR observers were 0.30 (s.e. = 0.04) and 0.35 (s.e. = 0.02) respectively. A 2x3 ANOVA with mode of viewing (BINOCULAR and MONOCULAR) as a between-subjects factor and image derivation (FD,ND and MIRROR) a within-subjects factor was used to analyse the hit rates. The mean hit rates (collapsed across image derivation) were 0.61 for binocular and 0.63 for monocular viewers. The ANOVA results showed no significant effect for mode of viewing ($F_{1,30} = 0.31; p > 0.5$). There was also no interaction between mode of viewing and image type ($F_{2,60} = 1.19, p > 0.25$). There was, however, a strong effect of image derivation ($F_{2,60} = 55.5, p < .0001$). The familiar direction, as expected, produced the best performance with mean hit rates equal to 0.84 (median=0.89, $d' = 1.92$) and 0.86 (median=0.89, $d' = 1.67$) for BINOCULAR and MONOCULAR observers respectively (see figure 9a). For the novel direction images, mean hit rates were equal to 0.54 (median=0.54, $d' = 0.72$) and 0.51 (median=0.54, $d' = 0.44$) for BINOCULAR and MONOCULAR viewing respectively. Figure 9a shows that only FD responses exceeded the 90% confidence threshold, as in the previous experiment, indicating once again that novel view generalization was much harder than in Experiment One. In a post hoc analysis we found only a significant difference between FD and the other two conditions ($p < .0001$). There was no significant difference between ND and MIRROR (for either mode of viewing; see figure 9a).

The correct response latencies are depicted in figure 9b. An ANOVA revealed no difference in response times for binocular and monocular viewing ($F_{1,30} = 2.02; p > .1$) nor an interaction between the two ($F_{2,60} = 0.06; p > .5$). There was a significant overall effect of image derivation ($F_{2,60} = 6.85; p < .0025$) with familiar views being recognized fastest in both groups.

In summary, this experiment revealed no differences in performance between monocularly and binocularly trained observers in terms of picture recognition although binocularly trained observers were more likely to identify the correct floorplan on their first test. There was a slight improvement in novel direction recognition performance compared to the previous experiment although a strict comparison is not possible because the current experiment used different target and distractor stimuli. Furthermore, sensitivity to novel views failed to reach the 90% confident limit, as in Experiment One, again indicating that generalisation to novel views is much more difficult without ego-motion.

5 Discussion

Our intention was to determine whether a generalisation to novel views is possible in the recognition of large scale spatial environments after restricted simulated movements through a scene during learning. We examine whether observers who learn to recognize a room from one specific set of directions can recognize the same room from different, unfamiliar, directions. We found in the first experiment that familiar and novel views were recognized, although familiar direction views were always recognized faster and more often. This is clearly an indication of view-dependency in internal representation immediately after training. The training used was both natural (in the sense that virtual head rotation and translation were possible) and extended (lasting approximately 20 minutes). An ego-centric encoding is also suggested by the fact that observers seldom considered the MIRROR condition images as familiar even though, in terms of features, they were identical to the familiar direction images. On the other

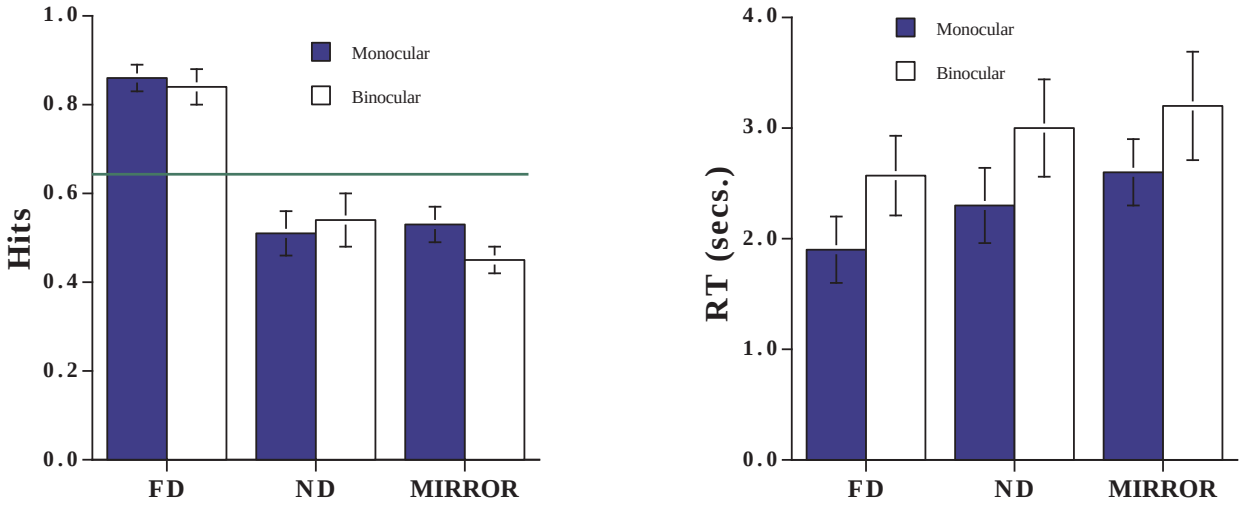

Figure 9: Proportion of hits (left) and responses times (right) for monocular and binocular observers as a function of image type for Exp. 3.

hand, the topographic floorplans were easily identified after training indicating that the encoding allows abstract structural detail to be reported.

However, these facts cannot readily be explained in terms of simple transformations on 2D stored-views since the novel views in our experiment were clearly distinct (in terms of heading) from the familiar views. The viewing restrictions applied meant that the minimum rotation in depth between views was 60° (maximum was 120°). These views contained features not visible from the familiar direction and lacked features that were (see for example figure 3). This makes it unlikely that observers identified novel views by simple image-based transformations.

In Experiments 2 and 3, in which passive viewing of images of the virtual room replaced active exploration, we found that this generalisation to novel views become more difficult to achieve. We attempted to preserve the same level of motivation by using a similar marker detection paradigm which also served to focus observers attention on the images they were to be tested on. The results for the two passive viewing experiments show that relatively few observers were able to recognize the novel views. Familiar view hit rates in both passive experiments was always above 75% showing that observers were able to recognize familiar directions but the reduced performance with novel views is again an indication that images such as these could not easily form the basis for novel direction view recognition. Something additional to 2D stored-views is necessary for generalisation to take place. It remains to be considered exactly what was afforded to the mobile observers and not to the passive observers. It is now generally accepted in computer vision that mobile or active observers can reduce the possibility of ambiguity and computational load by using a variety of strategies in the solution of specific tasks (Aloimonos, 1993; see also Blake & Yuille, 1992). In Experiment 3 we found no real benefit for observers trained using stereo images indicating that the disadvantage of passive observers was not the lack of depth cues afforded to mobile observers. However, in the detection of changes to spatial layout, Wang and Simons (1997) found a definite benefit for observers who were allowed to move. It could be that mobile observers can better learn the kinds of transformations that can occur in a given context and are therefore better able to compensate for novel views. This process would understandably be more computationally intensive (in that it requires extrapolation from familiar views). This would therefore explain the increases in response times for novel views. Furthermore the introduction of new structural detail into familiar views could account for the increased difficulty in making accurate familiarity judgements.

Finally, no mention has been made concerning

the benefits of the volitional, explorative, nature of the marker search for active observers. It is possible that the generalisation to novel views was made possible because the location of the 15 markers served as 3D landmarks and observers spent more time at these locations, adjusting their viewpoints, applying rotations of view etc. which produced benefits of the kind noted above. It remains to be seen what significance the volitional character of observer movement contributed to the novel view generalisation and also whether observers could indeed recognize novel views which have not had 'attentional' markers attributed to them. What is clear however is the importance of using natural yet controlled movement in general in studying the encoding of visual detail. Had we merely presented our observers with sequences of static images during their training then the ability to generalise to novel views may not have been noticed at all.

6 References

Aloimonos, Y. (1993) Active Perception, Hillsdale, NJ : Lawrence Erlbaum Associates.

Biederman I, 1987 Recognition-by-Components: A theory of human image understanding, Psychological Review, 94, 115-147

Biederman, I and Gerhardstein, PC (1992) Recognizing depth-rotated objects: Evidence for 3D viewpoint invariance, Journal of Experimental Psychology: Human perception and performance. 19(6), 1162-1182.

Blake,A. & Yuille, A. (1992) Active Vision, Cambridge, MA: MIT Press.

Bülthoff, HH and Edelman, S (1992) Psychophysical support for a 2-D view interpolation theory of object recognition, Proceedings of the National Academy of Science, 89, 60-64.

Bülthoff, HH, Edelman, S and Tarr, MJ (1994) How are three-dimensional objects represented in the brain? Technical Report No. 5, Max-Planck Institute for Biological Cybernetics, Tübingen.

Edelman, S and Bülthoff, HH (1992) Orientation dependence in the recogniton of familiar and novel views of three-dimensional objects, Vision Research, 32(12), 2385-2400.

Hock, HS and Schmelzkopf, KF (1980) The abstraction of schematic representations from photographs of real-world scenes, Memory and Cognition, 8(6) 543-554.

Macmillan,N.A. & Creelman C.D. (1991) Detection theory: A users guide, Cambridge,UK: Cambridge University Press.

Marr, D and Nishihara H.K. (1978) Representation and recognition of the spatial organization of three-dimensional shapes. Proc. R. Soc. London, 200, 269-294.

May, M Péruch P and Savoyant, A. (1995) Navigating in a virtual environment with map-acquired knowledge: encoding and alignment effects, Ecological Psychology, 7(1) 21-36.

Pinker, S. (1985) Visual Cognition: An introduction, in Visual Cognition S Pinker (ed.) Cambridge Mass.:MIT Press.

Rowland, GL, Franken, RE Bouchard, LM and Sookochoff, MB (1978) Recognition of familiar scenes from new perspectives, Perceptual and Motor Skills, 46(3,2), 1287-1292.

Shelton and McNamara (1997) Multiple views of spatial memory, Psychonomic Bulletin & Review, 4(1), 102-106.

Tarr, M and Pinker, S (1989) Mental rotation and orientation dependence in shape recognition, Cognitive Psychology, 21, 233-282.

Tarr, M.J. (1995) Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. Psychonomic Bulletin & Review, 2(1), 55-85.

Wang, RF and Simons, DJ (1997) Layout change detection is differentially affected by display rotations and observer movements, Investigative Ophthalmology and Visual Science, 38(4), 4695-B202.