

Fiona N. Newell · Andrew T. Woods ·  
Marion Mernagh · Heinrich H. Bülthoff

## Visual, haptic and crossmodal recognition of scenes

Received: 9 April 2004 / Accepted: 8 July 2004 / Published online: 15 October 2004  
© Springer-Verlag 2004

**Abstract** Real-world scene perception can often involve more than one sensory modality. Here we investigated the visual, haptic and crossmodal recognition of scenes of familiar objects. In three experiments participants first learned a scene of objects arranged in random positions on a platform. After learning, the experimenter swapped the position of two objects in the scene and the task for the participant was to identify the two swapped objects. In experiment 1, we found a cost in scene recognition performance when there was a change in sensory modality and scene orientation between learning and test. The cost in crossmodal performance was not due to the participants verbally encoding the objects (experiment 2) or by differences between serial and parallel encoding of the objects during haptic and visual learning, respectively (experiment 3). Instead, our findings suggest that differences between visual and haptic representations of space may affect the recognition of scenes of objects across these modalities.

**Keywords** Scene perception · Crossmodal recognition · Vision · Haptics · Orientation dependency

### Introduction

Our ability to recognise objects in a scene is not limited to the visual modality alone. We can, for example, recognise objects in a scene using touch without requiring visual input. We often search for an object in the absence of vision, such as looking for the flashlight in our tent at

night, using the relative positions and the identity of other objects that we feel as a guide to its possible location. The questions we address here are whether the same processes involved in visual scene recognition are also involved in haptic scene recognition, and whether or not these representations can be easily shared across sensory modalities.

Recent research has found that our visual memory for scenes is sensitive to the position of the observer (Christou and Bülthoff 1999; Diwadkar and McNamara 1997; Nakatani et al. 2002; Simons and Wang 1998; Wang and Simons 1999), suggesting that object relations are encoded in an egocentric, orientation-dependent manner. Recognition errors and latency typically increase with deviations in orientation between the learned and test view of the scene. Orientation-dependent performance disappears or is reduced, however, if the observer is allowed to move around a stationary display of the objects to observe a new view (Simons and Wang 1998; Simons et al. 2002; Wang and Simons 1999). Moreover, Simons et al. argue that this effect is not due to the visual background, although environmental cues can also help with spatial updating (Christou and Bülthoff 1999; McNamara 2003). Instead, their data suggest that extraretinal cues, such as proprioception or body movement, can help to spatially update the visual representation and thus compensate for subsequent changes in the view of the object or scene. From a stationary position, however, the evidence suggests that an egocentric representation of a visual scene dominates.

What is not yet known is whether the haptic representation of a scene is also dependent on the position of the observer during encoding. Studies that have investigated the manner in which objects or spatial locations are encoded by the haptic system provide inconsistent clues as to how scenes might be encoded. For example, Klatzky (1999) investigated haptic spatial representations using a path completion task. In her task, participants were first required to explore with their finger two sides of a triangle presented on a table top and then to complete the third side by returning to the origin. Klatzky

F. N. Newell (✉) · A. T. Woods · M. Mernagh  
Department of Psychology, University of Dublin,  
Trinity College,  
2 Dublin, Ireland  
e-mail: fiona.newell@tcd.ie  
Tel.: +353-1-6083914  
Fax: +353-1-6712006

H. H. Bülthoff  
Max-Planck Institute for Biological Cybernetics,  
Tübingen, Germany

measured angular and distance deviations as a function of the actual angle and distance to the origin. Furthermore, she tested effects of imagined changes in translation and rotation of the entire scene with respect to the participant on the observed responses. Klatzky reported that changes in imagined rotation resulted in more errors than changes in translation and that these errors were mainly angular response deviations rather than distance deviations. According to Klatzky, the effect of imagined changes in rotation were consistent with the adoption of a more object-centred representation of the paths. Other studies suggest that the representation of haptic spatial information is egocentric (Kappers 1999; Kappers and Koenderink 1999) although a recent study has suggested that haptic spatial representations can be more allocentric with delays between learning and test (Zuidhoek et al. 2003). We have recently reported evidence for an egocentric representation of single haptic objects (Newell et al. 2001). In a study into the effects of orientation on visual and haptic object recognition we found that recognition was orientation-dependent in both sensory modalities such that rotation changes of the objects by 180° between learning and test resulted in a cost in recognition performance. Scene perception, however, presents a different problem since there is an obvious increase in scale between single object recognition and scene recognition requiring large eye, arm and hand movements in order to encode the contents of the scene in vision and haptics. It is not clear from these studies, therefore, whether haptic scene recognition would adopt a more allocentric or object-centred representation of interobject relations or a more body-centred representation sensitive to deviations in orientation from the original learned position. We aim to elucidate the nature of scene representation in the haptic system and whether common processes such as viewer or body-centred representations exist for both vision and haptics.

A second aim of this paper is to investigate crossmodal efficiency between vision and haptics in the recognition of scenes. By investigating crossmodal recognition performance we would be able to provide a clearer understanding of the nature of the encoded information within each sensory modality and how this information is shared in order to recognise scenes of objects.

Recent research has shown that there are differences between the visual and haptic systems on how space is represented in memory. The perception of distance, for example, varies within and across these sensory modalities, and many studies have shown that the visual representation of horizontal and vertical distances are not equivalent (see, for example, Avery and Day 1969). In contrast, the haptic representation of distance in the radial and tangential dimensions is different (Marchetti and Lederman 1983) but there is no difference between the haptic representation of vertical and horizontal distance (Day and Avery 1970). Furthermore, adaptation to distances in one sensory modality does not transfer to the other modality suggesting that the representations of space are independent (Marks and Armstrong 1996). Further studies have shown that the nature of space

representation in touch is generally non-Euclidean (Blumenfeld 1937) resulting in distortions of distance estimation (Lederman et al. 1985) and judgements of parallelity across the workspace (Kappers 1999; Kappers and Koenderink 1999; Zuidhoek et al. 2003). Other studies have found large deviations in the haptic representation of the orientation of rods (Gentaz and Hatwell 1998, 1999) and paths (Faineteau et al. 2003). Together, these studies suggest that both the haptic and visual representation of spatial relations is distorted but in different ways across these sensory modalities. As such, matching representations of scenes across vision and haptics may not be optimal and may result in a cost in recognition performance.

Another difference between the visual and haptic senses is the encoding or exploratory procedures (Lederman and Klatzky 1987) adopted in scene perception. For example, it takes less than a second to capture the gist of a visual scene (Biederman et al. 1974; Potter 1976; Thorpe et al. 1996) suggesting that visual scenes are encoded in a holistic manner, although it is argued that such a representation may not be very detailed or complete (Rensink 2002; Rensink et al. 1997, 2000; Simons 1996). Haptic encoding, on the other hand, requires sampling the objects in a scene one at a time and a representation of the scene is subsequently built by integrating the interobject relations over time. Although the building of a rich representation of a visual scene may involve a similar temporal integration process across successive fixations in a visual scene (Brockmole et al. 2002), visual scene encoding can benefit from peripheral vision in a way that haptic encoding cannot. For example, Henderson and Hollingworth (2003) found that changes made to a peripheral target object in a scene could be detected more often than chance (see also Hollingworth et al. 2001), indicating that some aspects of peripheral objects are encoded prior to fixation. The number of objects that are actually encoded into a visual representation in memory is, however, a controversial issue. Some researchers suggest that the number of objects integrated in visual memory is limited to the number of objects that were attended (see Rensink 2002 for a review). Irwin and Zelinsky (2002) argued that the number of objects in a scene represented in visual memory is limited to about five items. More recently, Hollingworth (2003) has shown that the representations of objects in a visual scene can accumulate in visual memory due to orienting of the eyes and attention. In any case, it is clear that haptic encoding is limited to a localised, one-object-at-a-time encoding in a scene in order to build up a representation and does not benefit from parallel processing of the scene that may facilitate the subsequent representation of spatial configurations within a scene (Sanocki 2003).

Given these differences between vision and haptics, we tested whether scene representations are independent of sensory modalities or whether representations are modality specific. If a scene is represented independent of the encoding sensory modality then we would expect cross-modal recognition performance to be equivalent to within-

modality performance. On the other hand, if spatial relations between objects in a scene are represented in a manner specific to each sensory modality, then we would expect a cost in crossmodal recognition performance due to an incompatibility between the representations. We also tested whether orientation dependency was a process common to both sensory modalities. In experiment 1, we tested the visual, haptic and crossmodal recognition of scenes with changes in scene orientation. In experiments 2 and 3, we investigated possible encoding differences that might affect crossmodal scene perception.

## Experiment 1

### Materials and methods

#### Participants

Sixteen undergraduate students from the Eberhard-Karls University of Tübingen, Germany participated in this experiment for pay. Thirteen of the participants were female. Their ages ranged from 19 to 28 years old. All had normal or corrected-to-normal vision and none reported any haptic impairment. All participants gave written consent to take part in the experiment.<sup>1</sup>

#### Materials and apparatus

The stimulus set of objects included 15 wooden shapes of familiar objects. All objects were positioned on a rotating platform that was placed horizontally on top of a fixed table. The platform had 19 sunken position markers in which individual objects could be placed. Position markers were placed on the platform in such a way that each marker was equidistant from any of its neighbouring markers by a distance of 7 cm. The diameter of the platform measured 54 cm. When the participant was seated directly in front of the platform the furthest object position measured a distance of 42 cm and, therefore, all objects lay within manipulatory space (a term defined by Lederman et al. 1987). Furthermore, in this position all object markers were viewed from a slightly elevated angle and all object shapes were easily viewed without object occlusion.

The object stimuli were toy wooden shapes of animals painted white and placed on a narrow plastic stand that could be inserted in any position marker on the platform. All of these objects were 1 cm wide, varied in height from the base to the top of the object between 6 and 8 cm and between 3.5 and 5.5 cm in length.

In each trial, seven object stimuli were randomly chosen from the full set and placed in random positions on the platform. The stimuli themselves were placed in random

orientations with respect to direction in which the head of the animal pointed. The original orientation of the swapped objects was maintained between learning and test with respect to the scene. An example of the experimental set-up is illustrated in Fig. 1.

### Design

The experiment was based on a  $2 \times 2 \times 2$  factorial design using repeated measures. The main factors were the learning modality (vision or haptics), the sensory modality at test (within or across sensory modalities) and the change of orientation of the scene at test ( $0^\circ$  or  $60^\circ$ ). Trials were blocked into four different learning and testing blocks: Visual-Visual (VV), Haptic-Haptic (HH), Visual-Haptic (VH) and Haptic-Visual (HV). There were eight trials in each block, four at  $0^\circ$  orientation and four at  $60^\circ$  orientation. The order of the trials was random within blocks. Block order was counterbalanced across participants.

### Procedure

Each participant was seated in front of a table in such a way that their body midline was directly aligned with the centre of the scene. A trial consisted of scene learning followed by testing. During learning, participants were given 10 s to learn the scene visually or 60 s to learn the scene haptically. These timings were established in a pilot study and it allowed for equivalent performance across the visual and haptic sensory modalities. Participants were free to use their own exploration strategy but were



**Fig. 1** An illustration of our experimental set-up from a “bird’s eye” view. In this illustration the participant is engaging in a haptic scene recognition task

<sup>1</sup>All the studies reported were approved by the Trinity College Department of Psychology Ethics Committee, and thus conformed to the ethical standards laid down in the 1964 Declaration of Helsinki.

instructed not to touch the objects in the scene during visual learning. During haptic learning and testing, the participant placed their hands underneath a curtain to prevent viewing the scene, and could freely explore the objects using both hands. After visual learning, the experimenter lowered the curtain to cover the scene, and after haptic learning the participant was instructed to remove their hands from the objects. An interstimulus interval (ISI) of 20 s followed learning<sup>2</sup>, during which time the position of two of the seven objects was exchanged by the experimenter out of view. Furthermore, the orientation of the entire scene was either unchanged with respect to the observer, or rotated by 60°. We masked the sound of the platform rotating in all trials so that participants could not use secondary auditory cues to judge whether a change in orientation occurred during the ISI.

In the test, the participant had to identify which two of the seven objects were in new positions. Note that the global configuration of the objects in the scene was not affected by the exchange of the two objects because all original positions in the scene were still occupied. Scene recognition was tested either in the same sensory modality as learning or in the other modality. There was no time limit for the participant to make a response although they were prompted for an answer if none was forthcoming after a few minutes. Performance was measured in terms of error rates, in that, if the two swapped objects were correctly identified this was recorded as 0% error. If only one object was correctly identified this was recorded as a 50% error for that trial and if none of the objects were identified this was recorded as 100% error. The experiment began with a short practice block of four trials, one from each VV, HH, VH and HV block and two of which included a change in orientation. The experiment took approximately 80 min to complete.

## Results

The mean percentage error rates are presented in Fig. 2 across all conditions. We conducted a three-way ANOVA on the error data, with learning modality (vision or haptics), sensory modality at test (within or across) and orientation (0° or 60°) as factors. We found a main effect of sensory modality at test [ $F_{(1,15)}=9.0548$ ,  $MSE=249.27$ ,  $P<0.01$ ], with a greater number of errors occurring in the crossmodal than within-modal recognition conditions. We also found a main effect of orientation [ $F_{(1,15)}=19.722$ ,  $MSE=447.18$ ,  $P<0.001$ ], with more errors to 60° orientation change than 0° orientation change in the scene. There was no effect of learning modality [ $F_{(1,15)}=0.4267$ ,  $MSE=231.689$ ], and no interactions between the factors.

<sup>2</sup> An ISI of 20 s was necessary for the experimenter to move the objects. Although this ISI may contribute to a memory decay rate, a recent study suggests no difference in memory decay rates across vision and haptics (see Woods et al. 2004).

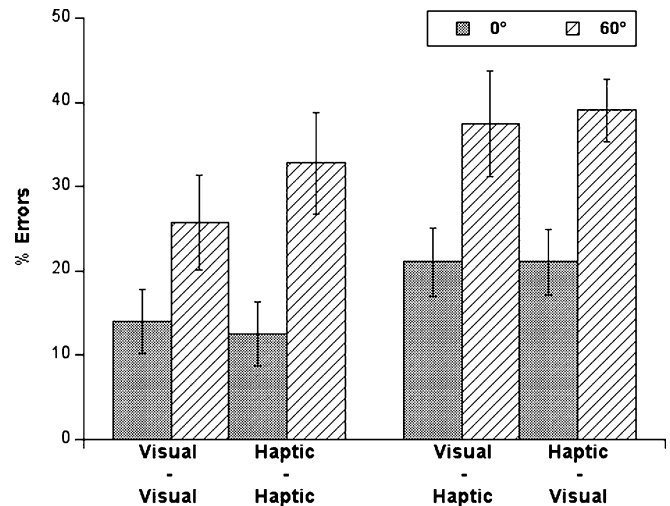


Fig. 2 Plot showing mean percentage errors made across all conditions in experiment 1. Error bars represent the standard error of the mean. The key indicates the orientation of the scene at test, i.e. either no change in orientation (0°) or the scene was rotated (60°)

## Discussion

We found that visual, haptic and crossmodal scene recognition was sensitive to the changes in the orientation of the scene with respect to the observer. More recognition errors occurred in both sensory modalities when there was a change in the orientation of the scene between learning and test. This result suggests that vision and haptics share a common, orientation-sensitive code when recognising scenes.

Our second main finding was that although recognition performance was always greater than chance, there was an overall cost in recognition when the test occurred across sensory modalities. Thus the encoded representation is not only specific to the orientation of the scene in each sensory modality but our data suggest that the representation is also specific to the encoding sense. This cost in recognition performance suggests that visual and haptic representations of spatial layout are not directly compatible and that a recoding is required in order to match one representation with the other. Such a recoding may be inefficient, resulting in errors in crossmodal scene recognition. Although our orientation-dependent result suggests that both representations are defined in egocentric coordinates (Newell et al. 2001), it does not necessarily suggest that the representations themselves are identical. As discussed earlier, one important difference between the sensory modalities is how space is represented. Previous studies suggest that haptic representation of space is distorted across the workspace (see, for example, Kappers 1999; Kappers and Koenderink 1999; Lederman et al. 1985). We suggest that it is the difference between the sensory modalities in how space is represented that affects the cost in crossmodal recognition.

There are, however, other differences between the sensory modalities in the encoding of the scenes that may explain the cost in crossmodal performance without

the need to appeal to representational differences. For example, it is possible that the objects in the scene were verbally recoded when learned through the visual or haptic systems. We used familiar objects as stimuli in our experiments. These objects were simple shapes of toy animals and are readily identifiable by sight and nameable after a few seconds of haptic palpation. Indeed, during the debriefing sessions our participants often referred to the objects by their name. In the following experiment we introduced an articulatory suppression task to control for the possibility that the objects were being recoded as lists of object names during visual or haptic learning. Another potential difference affecting crossmodal recognition is the nature of the encoding procedure between the senses. Whereas vision might encode important aspects of the scene in parallel, haptics encodes the constituent objects one at a time and a representation of spatial layout and object identities is integrated over time. In experiment 3 we attempt to control for these encoding differences across the senses.

---

## Experiment 2

In this experiment we investigated the role of verbal recoding in within and crossmodal scene recognition. A verbal mediation within or between vision and haptics would make our findings less generalisable, since it suggests that our test was not a direct test of visual, haptic or crossmodal memory. We were concerned that the objects used in our scenes were easily and readily labelled by their category names (for example, cow, dog, cat, etc.) when viewed. Furthermore, Millar (1975) found that verbal recoding improved tactile memory especially for similar tactile items, thus haptic performance may have been affected by verbal recoding. However, category naming takes longer by touch than by vision and often there may have been insufficient time to identify the objects by their category names. In our debriefing sessions, for example, participants often described the felt objects in terms of their physical features (for example, “vertical protrusion”, “bumpy top”, etc.) rather than their category names. The possibility remains, therefore, that if the stimuli were verbally recoded, better performance within sensory than across sensory modalities may have occurred because of the discrepancy between these descriptions across the sensory modalities.

There were two main aims to this experiment. The first was to test whether verbal recoding was involved in recognition performance for our scenes. The second was to determine if sensory modality-specific verbal recoding caused the cost in performance for crossmodal relative to within modal recognition. We used an articulatory suppression task during learning to test for effects of verbal recoding. If scenes of objects were recoded as verbal descriptions then we would expect a general cost in performance with articulatory suppression during learning. Furthermore, if verbal codes were modality specific, then we would expect that the cost in crossmodal relative to

within modal recognition would disappear with articulatory suppression.

## Materials and methods

### *Participants*

Twenty-four undergraduate and postgraduate students from the Department of Psychology, Trinity College Dublin participated in this experiment for research credits. Their ages ranged from 18 to 25 years old. Seventeen of the participants were female. All had normal or corrected-to-normal vision and none reported any haptic impairment. All participants gave written consent to take part in this experiment.

### *Stimuli and apparatus*

See experiment 1 for a description of the apparatus and stimuli used.

### *Design*

The experiment was based on a 2×2×2 repeated measures design with articulatory suppression (with or silent), learning modality (vision or haptics) and sensory modality at test (within or across) as factors. The experiment was divided into two main blocks, depending on whether articulatory suppression occurred during learning or not. The order of these blocks was counterbalanced across participants. Within each of these blocks there were four learning and test blocks (VV, HH, VH and HV), and four trials within each of these blocks. The order of the learning-test blocks was randomised across participants.

### *Procedure*

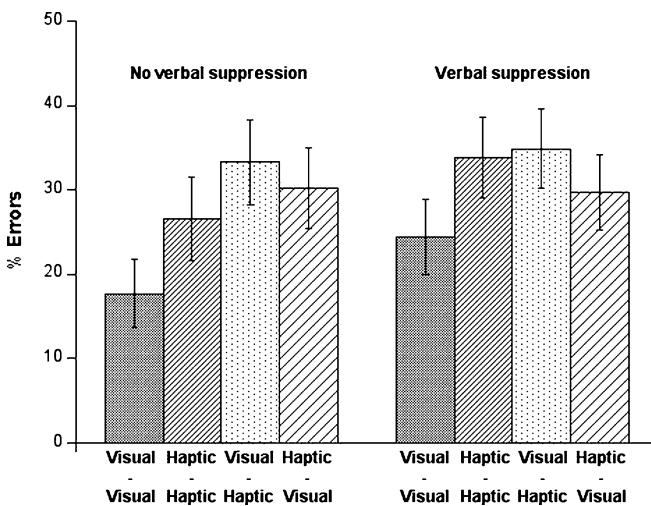
In this experiment there were two main experimental blocks, one was the same as the 0° trials in experiment 1, and the other was a repeat of these trials but participants were required to suppress any verbal encoding during learning. During the articulatory suppression block of the trials, participants were required to repeat aloud the word “the” at a rate of about 40 words per minute. This rate was demonstrated by the experimenter and practiced by the participant prior to the experiment. During the experiment, participants’ articulation was monitored by the experimenter to ensure that a steady rate was maintained during learning. This procedure ensured verbal encoding suppression without increasing cognitive demand (Baddeley 1986; Salamé and Baddeley 1982). All timings were the same as those in experiment 1.

## Results

The mean percentage error rates for each condition are plotted in Fig. 3. A three-way repeated measures ANOVA was conducted on the percentage correct responses with articulatory suppression (with or silent), learning modality (vision or haptics) and sensory modality at test (within or across) as factors. We found no effect of articulatory suppression [ $F_{(1,23)}=1.8319$ ,  $MSE=373.605$ ], and no effect of learning modality [ $F_{(1,23)}=0.6811$ ,  $MSE=431.35$ ]. A main effect of sensory modality at test was found [ $F_{(1,23)}=7.496$ ,  $MSE=260.66$ ,  $P<0.05$ ], with more errors made in the crossmodal conditions than the within modal conditions. There was a significant interaction between learning modality and sensory modality at test [ $F_{(1,23)}=14.274$ ,  $MSE=148.289$ ,  $P<0.001$ ]. A *post hoc* Newman-Keuls test on the interaction revealed that there were significantly fewer errors made in the VV condition than either the HH, VH or HV conditions ( $P<0.01$ ). The difference between VV and HH modalities was probably due to an increase in HH errors during the articulatory suppression task since there was no difference between VV and HH during the silent condition (Newman-Keuls,  $P>0.05$ ) but there was a difference between VV and HH with articulatory suppression (Newman-Keuls,  $P<0.05$ ). In other words, the number of errors made in the HH condition significantly increased relative to the VV condition with articulatory suppression. No other interactions between the factors were found.

## Discussion

In this experiment we found that articulatory suppression did not have an effect on scene recognition performance,



**Fig. 3** Plot showing mean percentage errors made across all conditions in experiment 2. Error bars represent the standard error of the mean. In the articulatory suppression task, participants were required to repeat aloud the word “the” at a minimum rate of 40 words per minute during visual and haptic learning. Otherwise the participants remained silent during learning (no articulatory suppression)

suggesting that a spatial representation of the scenes mediated recognition across sensory modalities rather than simply a verbal list of object names. Moreover, we replicated the cost of crossmodal recognition found in experiment 1. However, performance was worse for the within-modality haptic recognition (HH) of the scenes during articulatory suppression than for visual (VV) recognition.

The finding that articulatory suppression resulted in more errors in the HH condition than the VV condition might suggest a difference in encoding between these two sensory modalities. For example, the haptic representation may have some associated verbal coding which is vulnerable to a secondary articulatory suppression task. This suggestion is consistent with previous research where verbal interference was found to disrupt haptic memory (Mahrer and Miles 2002). However, we think it unlikely that haptic representations are always verbally recoded because this would predict that crossmodal recognition would be worse with articulatory suppression than without suppression. An alternative explanation may be that verbal recoding was used as a response strategy during the HH block only. In any case, there clearly is sufficient shape-specific information in the haptic representation to allow for crossmodal recognition irrespective of the learning modality. In summary, scene recognition performance is not impaired by articulatory suppression and, therefore, crossmodal recognition is unlikely to be mediated by verbal recoding.

## Experiment 3

Differences in encoding procedures between the sensory modalities may account for the cost in crossmodal scene recognition. For example, when a participant was presented with a scene of objects that they had to learn through vision, it is conceivable that elements of the scene were rapidly encoded in parallel which could subsequently benefit scene recognition (Biederman et al. 1974; Sanocki 2003; Sanocki and Epstein 1997). On the other hand, haptic learning was necessarily serial because the hands moved across the scene one object at a time. In this case, the relative location of the objects with respect to one another had to be derived over time. These encoding differences may have affected a cost in crossmodal recognition if the encoded representation from one sensory modality would not directly match the representation of the other. In this experiment we tried to reduce the differences in encoding the scenes of objects between the sensory modalities by presenting the objects serially during visual learning. We adopted a procedure used in previous studies comparing haptic and visual picture perception (Loomis et al. 1991) that allowed for partial viewing of the contents of the scene during visual learning. We predicted that if the cost in crossmodal recognition was based on serial versus holistic encoding differences between haptics and vision, respectively, then

this cost should disappear when encoding conditions were equivalent across the sensory modalities.

## Materials and methods

### Participants

Sixteen undergraduate students from the Department of Psychology, Trinity College Dublin participated in this experiment for research credits. Twelve of the participants were female. Their ages ranged from 18 to 22 years old. None of these volunteers had participated in any of the previous experiments. All participants had normal or corrected-to-normal vision and none reported any haptic impairments. Each participant gave written consent to take part in the experiment.

### Stimuli and apparatus

The stimuli and apparatus were the same as those described in experiment 1 with the following exceptions. We used seven black cylinders which were open at one end and when upturned fitted perfectly over each object in a scene. These cylinders were used to cover the objects only during visual learning of the scene.

### Design

The experiment was based on a within subject,  $2 \times 2$  design with sensory modality at test (within or across) and learning modality (vision or haptics) as factors. There were four trials within each of the VV, HH, VH and HV testing blocks. The order of the trials was randomised within a block, and the order of the blocks was counterbalanced across participants.

### Procedure

The procedure followed that outlined in experiment 1 with the following exceptions. Here participants were given 15 s to learn the scene visually which allowed extra time to uncover each of the objects in the scene. During visual learning, the scene was presented to the participant but each object on the scene was covered. The participant was then required to lift the cover to reveal the object underneath and to repeat this procedure for each object on the scene. Haptic learning conditions and all test conditions were the same as in experiment 1. We did not test effects of orientation in this experiment. The experiment took approximately 40 min to complete.

## Results

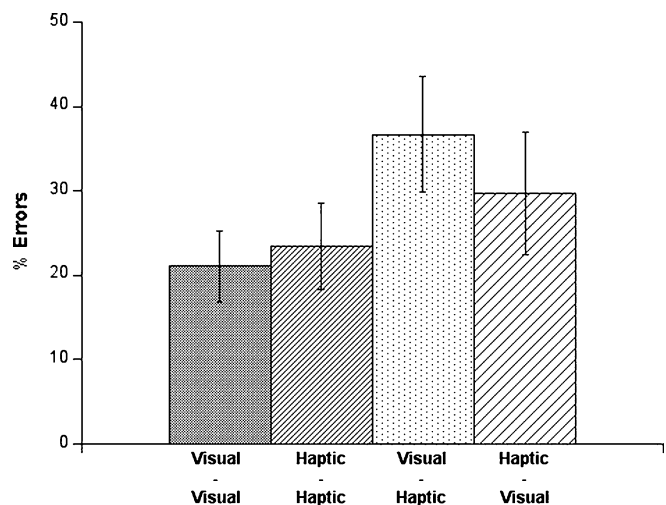
The mean percentage error rates across each of the test conditions are shown in Fig. 4. We conducted a two-way, repeated measures ANOVA on the error data with learning modality (vision or haptics) and sensory modality at test (within or across) as factors. We found a main effect of sensory modality at test [ $F_{(1,15)}=6.176$ ,  $MSE=309.895$ ,  $P<0.05$ ], with a greater number of errors made in the crossmodal test conditions than the within modal conditions. There was no effect of learning modality [ $F_{(1,15)}=0.14438$ ,  $MSE=608.723$ ], and no interaction between the factors [ $F_{(1,15)}=1.421$ ,  $MSE=247.395$ ].

## Discussion

We were concerned that the cost in crossmodal recognition found in our previous experiments was based on the differences in encoding the objects across sensory modalities. In particular, vision allows for more holistic encoding of a scene whereas with haptics, the objects in a scene are learned one at a time. In this experiment, participants learned the objects in a scene one at a time for both vision and haptics. We found that despite this equivalent encoding procedure, the cost in crossmodal recognition performance remained. This cost, therefore, cannot be attributed to differences in serial versus holistic encoding of scenes between the two modalities.

## General discussion

In our study we replicated previous studies reporting orientation dependency in the visual recognition of scenes (Christou and Bühlhoff 1999; Diwadkar and McNamara 1997; Nakatani et al. 2002; Simons and Wang 1998; Wang



**Fig. 4** Plot showing mean percentage errors made across all conditions in experiment 3. Error bars represent the standard error of the mean. In this experiment participants were required to learn the objects one at a time during both visual and haptic learning

and Simons 1999) and we present a new finding that haptic scene perception is also orientation dependent. For both visual and haptic recognition, performance decreased when the scene was rotated by 60° with respect to the observer. This finding suggests that the mechanisms involved in recognising scenes are the same for vision and haptics. At least, when the observer is stationary the spatial layout of the objects in a scene is represented in terms of an egocentric reference frame. These initial investigations suggest a functional similarity between vision and haptics in how scenes of objects are represented. We will return to this point later.

Although the scenes were encoded with respect to the position of the observer in both sensory modalities, sufficient differences existed between the visual and haptic representations to cause a decrease in recognition performance with a change in modality between learning and test. In experiments 2 and 3, we ruled out encoding differences between the modalities, such as explicit object naming and serial or holistic encoding, which may have affected this cost in crossmodal performance. It is not clear from our studies what contributes to the cost in crossmodal performance, even when there is no change in orientation of the scenes. This finding suggests that scene representation is sensory modality specific but does not explain the precise nature of the representations of scenes within each sensory modality. Although such an investigation is beyond the scope of the current study we can nevertheless offer several reasons that may explain the cost in recognition that occurs with a change of sensory modality.

First, it might be that the cost in crossmodal recognition is due to differences in the processes involved in encoding scenes across the sensory modalities. Some recent studies have provided evidence for two processes in visual scene perception: an encoding of object positions and an encoding of object identity relative to each position (Kosslyn et al. 1992; Postma and de Haan 1996; Sanocki 2003). Other studies have highlighted the relationship between these two processes. For example, Aginsky and Tarr (2000) argued that the position and presence of objects are readily encoded in a scene and that these properties may help determine the configuration of the exact objects in a scene. Similarly, Sanocki and colleagues (Sanocki 2003; Sanocki and Epstein 1997) reported that the representation of scene layout could facilitate the subsequent perception of the spatial relations between objects within a scene. It is not obvious that haptics could also involve two processes as in vision. It is conceivable that during haptic learning of a scene the hands would quickly sweep over the entire scene to encode the position and presence of objects. We did not observe this, although the nature of our task may have rendered such a rapid acquisition of scene layout redundant as only object-to-position information was required to perform the task and the global configuration of the scene did not change between learning and test. If global configuration did change then we may have observed a rapid scanning by the hands of the scene layout prior to object-by-object exploration. Since the layout of the object positions was

non-informative for the purposes of the task, the rapid encoding of scene layout may not have benefited visual scene representation relative to haptic representation.

Another possible reason why a cost in crossmodal recognition occurred may be because of differences in the nature of the object attributes that were encoded across the modalities. For example, Lederman et al. (1996) have argued that certain features of stimuli can result in a modality encoding bias resulting in a rich representation in one modality relative to the other and consequently poor crossmodal performance. (An extreme example of this is that colour can only be encoded by vision and if stimuli differed only along the colour dimension then crossmodal performance would be impaired.) Similarly, it is argued that vision allows for encoding of each object shape in a holistic manner (see, for example, Tarr and Bülthoff 1998) whereas haptics may not allow for rapid holistic encoding of shape. Our scenes were relatively complex and, therefore, when the objects are explored by touch it is possible that a piecemeal representation of the scene was built based on the salient features of each object and not the entire shape of the objects. The encoded features for optimal visual performance, therefore, might not be the same as those encoded for optimal haptic performance, resulting in poor feature matching across but not within sensory modalities. However, in our experiments here we think it unlikely that a modality encoding bias occurred because we made every effort to ensure that all features of our stimuli were accessible to both modalities and that no feature could be uniquely encoded by one modality only. We also think it unlikely that only the salient object properties were encoded by haptics since in all cases the participants used a contour-tracing exploratory procedure on each object during haptic learning (Lederman and Klatzky 1987) suggesting that the entire object shape was encoded. Furthermore, participants were aware which learning and testing block they were about to embark on and it would not be clear why a participant would adapt a learning procedure that would inevitably result in poor performance. Nevertheless, time constraints may have prevented rich haptic representations of the objects to be stored and, therefore, it is possible that with changes in the task, such as an increase in learning time or using less complex shapes, more efficient crossmodal recognition performance would be evident.

We propose that the most likely reason why crossmodal performance was less efficient than within modal performance is because of differences in how spatial layout is represented across vision and haptics. This proposal is consistent with previous literature reporting systematic distortions in the haptic representation of space compared to the visual representation of space (Blumenfeld 1937; Kappers 1999; Kappers and Koenderink 1999; Marks and Armstrong 1996). In particular, Kappers noted that the representation of haptic space becomes systematically more compressed away from the axis aligned to the body midline. If such distortions of haptic representational space are indeed contributing to the cost in crossmodal performance then we would predict that a

shift of the centre of the scene away from the body midline would result in a further decrease in crossmodal performance. We are currently embarking on a series of studies designed to investigate the role of spatial representations across vision and haptics in crossmodal recognition.

Finally, our finding that both the visual and haptic modalities encode scenes of objects in relation to a body-centred or ego-centred reference frame is consistent with the literature on reference frames in vision and touch. McNamara (2003) recently argued that the layout of visual scenes is interpreted in terms of a spatial reference system intrinsic to the set of objects in the scene. The reference system, therefore, is either egocentric or environment-centred depending on available cues and the layout of the objects themselves. For example, the specific arrangement of objects in a scene of objects (for example, arranged in rows) or the orientation of the room in which the objects are placed provide salient environmental reference frames by which these objects can be encoded. McNamara, however, argues that the dominant reference frame by which scenes are generally encoded by the visual system is the egocentric reference frame. Interestingly, Newport et al. (2002) suggest that visual reference frames can also influence haptic perception of space. In their study participants were required to match the rotation of a bar presented on a table top to the orientation of an adjacent reference bar so that they were parallel to each other. For one half of the trials the participants were blindfolded while they performed the task. In the other half of the trials the participants were not blindfolded but the stimuli were placed underneath a wooden board out of sight. Newport et al. replicated the effect of large deviations in parallelity when the participants were blindfolded (see also Kappers 1999). However, when participants were not blindfolded accuracy levels significantly increased, suggesting that non-informative visual cues aided in the haptic representation of space. It is possible that under our experimental conditions non-informative visual information and context, such as the layout of the room, or the table on which the scene platform was placed, offered a reference frame by which the haptic scene was encoded. Since performance was sensitive to changes in the orientation of the scene we can conclude that an egocentric reference frame was adopted by both sensory modalities, although visual cues may have had an effect on the haptic modality. An egocentric reference frame might not have been adopted by the haptic system if our participants had been blindfolded, and we might expect haptic scene perception to be worse in this case than when participants were able to view the surrounding environment. Clearly further research is required in order to provide a better understanding of how spatial configurations of objects in our world are represented within and across our senses.

**Acknowledgements** This work was funded by a European Union IST programme grant (IST-2001-34712) awarded to the first author and by the Max-Planck Society, Germany. We thank Karl-Heinz Hofmann and Christina Baum of the Max-Planck Institute for Biological Cybernetics for building our experimental apparatus.

## References

- Aginsky V, Tarr MJ (2000) How are different properties of a scene encoded in visual memory? *Vis Cogn* 7:147–162
- Avery GC, Day RH (1969) Basis of horizontal-vertical illusion. *J Exp Psychol* 81:376–380
- Baddeley AD (1986) Working memory. Oxford University Press, Oxford, UK
- Biederman I, Rabinowitz JC, Glass AL, Stacey EWJ (1974) On the information extracted from a glance at a scene. *J Exp Psychol* 103:597–600
- Blumenfeld W (1937) The relationship between optical and haptic construction of space. *Acta Psychol* 2:125–175
- Brockmole JR, Wang RF, Irwin DE (2002) Temporal integration between visual images and visual percepts. *J Exp Psychol Hum Percept Perform* 28:315–334
- Christou CG, Bühlhoff HH (1999) View dependence in scene recognition after active learning. *Mem Cogn* 27:996–1007
- Day RH, Avery GC (1970) Absence of the horizontal-vertical illusion in haptic space. *J Exp Psychol Gen* 83:172–173
- Diwadkar VA, McNamara TP (1997) Viewpoint dependence in scene recognition. *Psychol Sci* 8:302–307
- Faineteau H, Gentaz E, Viviani P (2003) The kinaesthetic perception of Euclidean distance: a study of the detour effect. *Exp Brain Res* 152:166–172
- Gentaz E, Hatwell Y (1998) The haptic oblique effect in the perception of rod orientation by blind adults. *Percept Psychophys* 60:157–167
- Gentaz E, Hatwell Y (1999) Role of memorization conditions in the haptic processing of orientations and the ‘oblique effect’. *Br J Psychol* 90:373–388
- Henderson JM, Hollingworth A (2003) Eye movements and visual memory: detecting changes to saccade targets in scenes. *Percept Psychophys* 65:58–71
- Hollingworth A (2003) Failures of retrieval and comparison constrain change detection in natural scenes. *J Exp Psychol* 29:388–403
- Hollingworth A, Williams CC, Henderson JM (2001) To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonom Bull Rev* 8:761–768
- Irwin DE, Zelinsky GJ (2002) Eye movements and scene perception: memory for things observed. *Percept Psychophys* 64:822–895
- Kappers AM (1999) Large systematic deviations in the haptic perception of parallelity. *Perception* 28:1001–1012
- Kappers AM, Koenderink JJ (1999) Haptic perception of spatial relations. *Perception* 28:781–795
- Klatzky RL (1999) Path completion after haptic exploration without vision: implications for haptic spatial representations. *Percept Psychophys* 61:220–235
- Kosslyn SM, Chabris CF, Marsolek CJ, Koenig O (1992) Categorical versus coordinate spatial relations: computational analyses and computer simulations. *J Exp Psychol Hum Percept Perform* 18:562–577
- Lederman SJ, Klatzky RL (1987) Hand movements: a window into haptic object recognition. *Cogn Psychol* 19:342–368
- Lederman SJ, Klatzky RL, Barber PO (1985) Spatial and movement-based heuristics for encoding pattern information through touch. *J Exp Psychol Gen* 114:33–49
- Lederman SJ, Klatzky RL, Collins A, Wardell J (1987) Exploring environments by hand or foot: time-based heuristics for encoding distance in movement space. *J Exp Psychol Learn Mem Cogn* 13:606–614

- Lederman SJ, Summers C, Klatzky RL (1996) Cognitive salience of haptic object properties: role of modality-encoding bias. *Perception* 25:983–998
- Loomis JM, Klatzky RL, Lederman SJ (1991) Similarity of tactual and visual picture recognition with limited field of view. *Perception* 20:167–177
- Mahrer P, Miles C (2002) Recognition memory for tactile sequences. *Memory* 10:7–20
- Marchetti FM, Lederman SJ (1983) The haptic radial-tangential effect: two sets of Wong's (1977) "moments-of-inertia" hypothesis. *Bull Psychonom Soc* 21:43–46
- Marks LE, Armstrong L (1996) Haptic and visual representations of space. In: Inui T, McClelland JL (eds) *Attention and performance, XVI. Information integration in perception and communication*. MIT Press, Cambridge, MA, pp 263–287
- McNamara TP (2003) How are locations of objects in the environment represented in memory? In: Freska C, Brauer W, Habel C, Wender K (eds) *Spatial cognition, III. Routes and navigation, human memory and learning, spatial representation and spatial reasoning*. Springer, Berlin Heidelberg New York, pp 174–191
- Millar S (1975) Effects of tactual and phonological similarity on the recall of Braille letters by blind children. *Br J Psychol* 66:193–201
- Nakatani C, Pollatsek A, Johnson SH (2002) Viewpoint-dependent recognition of scenes. *Q J Exp Psychol* 55A:115–139
- Newell FN, Ernst MO, Tjan BS, Bühlhoff HH (2001) Viewpoint dependence in visual and haptic object recognition. *Psychol Sci* 12:37–42
- Newport R, Rabb B, Jackson SR (2002) Noninformative vision improves haptic spatial perception. *Curr Biol* 12:1661–1664
- Postma A, de Haan EHF (1996) What was where? Memory for object locations. *Q J Exp Psychol* 49A:187–199
- Potter MC (1976) Short-term conceptual memory for pictures. *J Exp Psychol Hum Learn Mem* 2:509–522
- Rensink RA (2002) Change detection. *Annu Rev Psychol* 53:245–277
- Rensink RA, O'Regan JK, Clark JJ (1997) To see or not to see: the need for attention to perceive changes in scenes. *Psychol Sci* 8:368–373
- Rensink RA, O'Regan JK, Clark JJ (2000) On the failure to detect changes in scenes across brief interruptions. *Vis Cogn* 7:127–145
- Salamé P, Baddeley AD (1982) Disruption of short-term memory by irrelevant speech: implications for the structure of working memory. *J Verbal Learn Verbal Behav* 21:150–164
- Sanocki T (2003) Representation and perception of scenic layout. *Cogn Psychol* 47:43–86
- Sanocki T, Epstein W (1997) Priming spatial layout of scenes. *Psychol Sci* 8:374–378
- Simons DJ (1996) In sight, out of mind: when object representations fail. *Psychol Sci* 7:301–305
- Simons DJ, Wang RF (1998) Perceiving real-world viewpoint changes. *Psychol Sci* 9:315–320
- Simons DJ, Wang RF, Roddenberry D (2002) Object recognition is mediated by extraretinal information. *Percept Psychophys* 64:521–530
- Tarr MJ, Bühlhoff HH (1998) Image-based object recognition in man, monkey and machine. *Cognition* 67:1–20
- Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. *Nature* 381:520–522
- Wang RF, Simons DJ (1999) Active and passive scene recognition across views. *Cognition* 70:191–210
- Woods AT, O'Modhrain S, Newell FN (2004) The effect of temporal delay and spatial differences on crossmodal object recognition. *Cogn Affective Behav Neurosci* (in press)
- Zuidhoek S, Kappers AML, van der Lubbe RHJ, Postma A (2003) Delay improves performance on a haptic spatial matching task. *Exp Brain Res* 149:320–330