Cross-modal perception of actively explored objects

Fiona N.Newell, Heinrich H. Bülthoff and Marc O. Ernst

Department of Psychology, Trinity College Dublin, Ireland.
fiona.newell@tcd.ie
Max-Planck Institute for Biological Cybernetics, Tübingen, Germany.
{heinrich.buelthoff, marc.ernst}@tuebingen.mpg.de

Abstract. Many objects in our world can be picked up and freely manipulated, thus allowing information about an object to be available to both the visual and haptic systems. However, we understand very little about how object information is shared across the modalities. Under constrained viewing cross-modal object recognition is most efficient when the same surface of an object is presented to the visual and haptic systems [5]. Here we tested cross modal recognition of novel objects under active manipulation and unconstrained viewing of the objects. These objects were designed such that each surface of the object provided unique information. In Experiment 1, participants were allowed 30 seconds to learn the objects visually or haptically. Haptic learning resulted in relatively poor haptic recognition performance relative to visual recognition. In Experiment 2, we increased the learning time for haptic exploration and found equivalent haptic and visual recognition, but a cost in cross modal recognition. In Experiment 3, participants learned the objects using both modalities together, vision alone or haptics alone. Recognition performance was tested using both modalities together. We found that recognition performance was significantly better when objects were learned by both modalities than either of the modalities alone. Our results suggest that efficient cross modal performance depends on the spatial correspondence of object surface information across modalities.

1 Introduction

Information about objects in our world can be gathered using many senses. In order to recognise objects, spatial properties such as form and surface texture can be determined by our visual and haptic systems. By combining information across the senses recognition can be more efficient. For example, it is well documented that changes in viewpoint of an object with respect to the observer can decrease recognition performance [e.g. 1, 2, 3, 4, 6, 7]. However, by combining information about surface properties of objects across vision and haptics, the effect of viewpoint is reduced [5]. In our previous studies we found that the back surface of an object, which was fixed in space, was better represented by the haptic system whereas the front of the object was better represented by the visual system. By combining information about the object across modalities a richer object representation is created which allows for more view independent recognition.

We found that surfaces information of objects that are fixed in space can be combined across the modalities to result in more efficient recognition. The question then arises as to whether or not information about surfaces of objects that can be freely explored is also combined across the senses, particularly when each surface of the object is unique. Sensory combination may require spatial constraints to allow for efficient surface matching across the modalities. If objects are fixed in space, then the fewer surfaces encoded by vision or touch, the more likely a match can be made across modalities. However, if objects are freely explored the possible number of surfaces encoded by each modality is considerably larger, therefore matching may become more difficult across the modalities. On the other hand, if objects are freely explored by both modalities simultaneously, then correspondence across the surfaces is directly encoded and should lead to better recognition performance. These issues are explored in the following series of experiments.

2 Experiment 1

We tested observer sensitivity to spatial changes when two "L-shaped" objects were presented 0s, 15s or 30s apart by measuring accuracy. Reaction times (RT's) were also measured.

2.1 Method

24 members of the Max Planck Institute for Biological Cybernetics participated in this experiment. Seven of he participants were female. Their ages ranged from 27 to 39 years. All participants were naïve to the purposes of the task and all had normal or corrected-to-normal vision.

We used the same stimulus set of objects as described in Newell et al. (2001): All objects were composed of 6, same sized Lego bricks (measuring 3x1.5 cm). Each object was defined by a unique configuration of these bricks (see Figure 1 for an illustration). All of the bricks were red in colour, therefore, we eliminated the possibility of a modality encoding bias due to changes in weight, size or colour between the objects. We created 32 individual object stimuli for our experiment and an extra 4 objects for practice.

For the visual conditions, each object was placed inside a transparent perspex sphere, the kind often used for homemade Christmas tree decorations. The diameter of a sphere was cm, therefore an object fitted neatly inside. By using the sphere, participants could freely view the object without touching it.

The experiment was based on a two-factor repeated measures design with learning modality (vision or haptics) and modality conditions (within or across modality) as factors. The experiment was divided into four separate blocks with different learning and testing conditions, two within-modality; visual-visual (V-V) and haptic-haptic (H-H), and two across modalities; visual-haptic (V-H) and haptic-visual (H-V). The order of these blocks was counter-balanced across participants. For each participant, the 32 objects were randomly assigned to each experimental block and each object was randomly assigned as either a target or a non-target within each block.

The experiment was divided into four separate experimental blocks and participants could take a self-timed break between each block. Within each block the participant was first required to learn four target objects. During learning, the target objects were presented in sequence and the participant was given 30 seconds to learn each object, irrespective of modality. Their recognition of the targets was subsequently tested using an old/new recognition memory protocol. That is, participants were presented with objects, one at a time, and asked whether each object was a target or not. The four target objects and four new, non-target objects were presented in a random order within each block. Thus there were 8 experimental trials per block. We also repeated 4 trials at random within each block in order to avoid participants using a 50% response guessing strategy but we did not record responses to these trials. The experiment was preceded by four practice trials, one from each block.

During learning and testing the participant could freely view or palpate the objects depending on the experimental condition. Participants placed their hands underneath a curtain screen during haptic exploration, therefore the object was completely out of sight. For visual exploration, the objects were placed inside a clear sphere which was placed in the hands of the participant. Thus all surfaces of the object could be freely viewed but could not be touched. Participants were instructed as to the nature of the learning and test modality prior to each experimental block. The participants' responses were recorded on a scoring sheet by the experimenter. The experiment took approximately one hour to complete.

2.2 Results and Discussion

The mean percent correct scores (hits and correct rejections) are plotted in Figure 1. A two way ANOVA was conducted on the mean number of correct responses across the learning and recognition conditions. We found no main effect of modality [$F(1,23)=2.195$, $p=0.1520$]. A main effect of learning modality was found, [$F(1,23)=10.84$, $p<0.005$]. Recognition was better when the objects were learned visually than haptically. There was no interaction between the factors [$F(1,23)=0.4753$, n.s.]. An analysis of the hit trials only (i.e. correct targets identified) showed the same pattern of results; no effect of transfer condition [$F(1,23)<1$], a main effect of learning [$F(1,23)=6.053$, $p<0.05$] and no interaction [$F(1,23)=2.189$, n.s.].

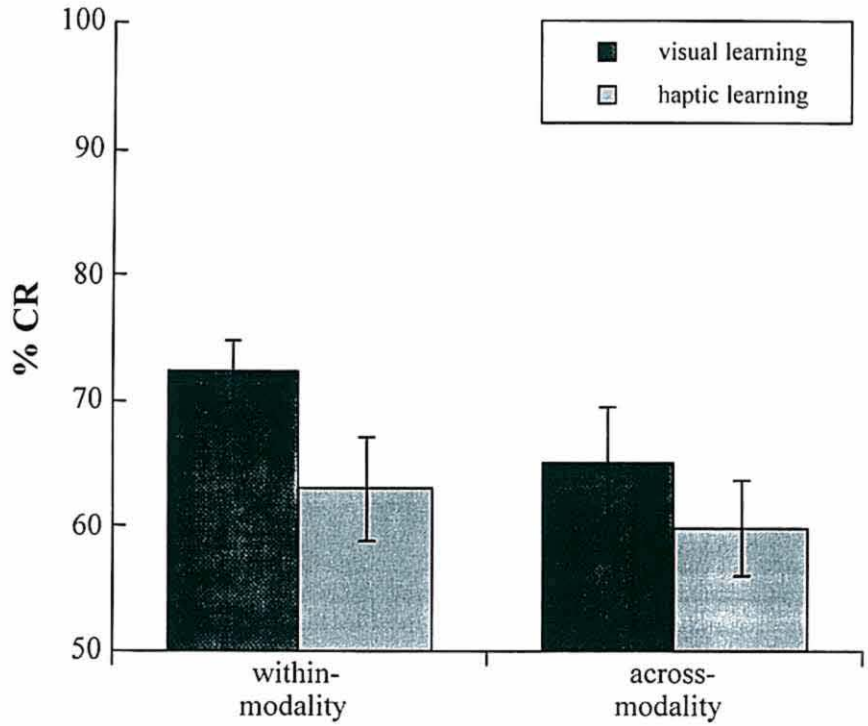

Fig. 1. Plot showing mean correct responses in each of the recognition conditions. Here, within-modality visual recognition was better than performance in all other conditions.

We found that objects which were learned visually were easier to recognise than objects learned haptically. Although there was no indication that recognition performance was better within modalities than across, any effect may have been obscured by the relatively better visual learning. Participants were given the same amount of time to learn the objects either haptically or visually. Given that haptic information pick-up may be slower than visual encoding, we decided to increase learning time for haptics in the following experiment.

3 Experiment 2

In this experiment, we were interested in measuring the effects of cross-modal recognition when encoding was equivalent across the visual and haptic modalities.

Here we replicated Experiment 1 with the exception that participants were now given 60 seconds to learn the target objects in the haptic condition.

3.1 Method

18 individuals were recruited from the Max Planck Institute for Biological Cybernetics Subject List to participate in the experiment for pay (8 euro per hour). Five of the participants were female. Their ages ranged from 22 to 36 years. All participants were naïve to the purposes of the task and all had normal, or corrected to normal, vision.

See Experiment 1 for a description of the stimuli used.

The experiment followed the same design used in Experiment 1. In this experiment the procedure was slightly different: Participants were given 30 seconds to learn the objects visually (as in the previous experiment) but was increased to 60 seconds in the haptic condition.

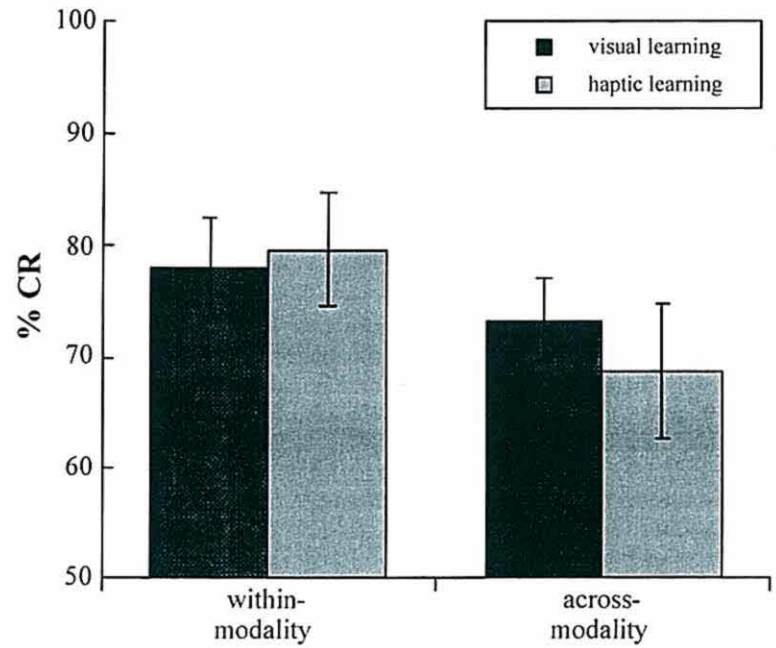

Fig. 1. Plot showing mean percent correct scores across the learning and recognition conditions in Experiment 2. Within modality recognition performance was significantly better than cross-modality performance.

3.2 Results and Discussion

The mean percent correct scores (hits and correct rejections) are plotted in Figure 2. A two way ANOVA was conducted on the mean number of correct responses across the learning and recognition modalities (see Figure 2). We found a main effect of recognition modality condition [$F(1,15)=7.653$, $p<0.05$], indicating that performance was better for within modal recognition than across modalities. There was no main effect of learning modality, [$F(1,23)=0.1094$, n.s.] and no interaction between the factors [$F(1,15)=0.909$, n.s.]. An analysis of the hit trials only (i.e. correct targets identified) showed the same pattern of results; a main effect of recognition condition [$F(1,15)=10.075$, $p<0.01$], no effect of learning [$F(1,15)<1$] and no interaction [$F(1,15)<1$].

When we increased the learning time in the haptic condition from 30 seconds (in Experiment 1) to 60 seconds in this experiment we found significantly better within-modality recognition performance than cross-modal performance. In this experiment, objects were presented to one modality at a time. In the following experiment we test the effect on recognition performance of presenting objects to both modalities simultaneously.

4 Experiment 3

In our previous studies (see Newell et al. 2001) we found that cross modal performance was the same as within modal performance when there was a change in the orientation of the object. We argued that visual and haptic modalities correspond with each other based on shared information about surface of objects. Thus, performance was best when the visual system viewed a particular surface of an object from the front and the haptic system was presented with the same surface of the object from the back. The spatial correspondence of object properties across modalities seems, therefore, to be important for cross-modal perception.

We might argue that the results from Experiment 2 are due to poor spatial correspondence between vision and haptics. When objects are freely manipulated or actively viewed, information from all surfaces of the objects can be learned but exploration is necessarily unconstrained. This may provide circumstances where spatial correspondence across the modalities is poor because matching the object surfaces across modalities may be more difficult. If, on the other hand, objects were freely palpated in the presence of vision, then spatial correspondence would be direct and should promote better recognition performance.

To test this idea, we allowed participants to learn objects either using vision alone, haptics alone or bimodally (i.e. using touch and vision together). We predicted that bimodal learning would result in better cross-modal recognition than uni-modal learning, because information about the objects' surfaces was encoded by both haptics and vision directly and in correspondence.

4.1 Method

24 persons were recruited from the Max Planck Institute for Biological Cybernetics Subject List to participate in the experiment for pay (8 euro per hour). All participants were naïve to the purposes of the task and all had normal, or corrected to normal, vision.

24 objects were used as stimuli. See Experiment 1 for a general description of the objects.

The experiment was based on a one-way, within subjects design with learning modality as the factor (visual, haptic or bimodal). Recognition was always tested bimodally, i.e. using vision and haptics together. The experiment was divided into three separate blocks, each with a different learning condition, (i.e. V-VH, H-VH and VH-VH representing the learning and testing modalities respectively). The order of these blocks was counter-balanced across participants. For each participant, the 24 objects were randomly assigned to each experimental block and each object was randomly assigned as either a target or a non-target within each block.

The procedure mainly followed that outlined in Experiment 1 with the following exceptions: There were three experimental blocks and participants could take a self-timed break between each block. The blocks were differentiated by the learning modality, i.e. whether learning was visual, haptic or bimodal. For visual learning, the objects were placed inside a perspex sphere for viewing. During haptic learning the objects were placed behind a curtain under which the participant placed their hands. During bimodal learning the participant could freely palpate and view the objects simultaneously. Participants were given 60 seconds to learn the target objects in each learning condition. There were 4 target and 4 non-target objects in each block. The order of the blocks was counter-balanced across participants. Testing was always conducted bimodally. Again we ran 12 trials per block in order to avoid a guessing strategy.

4.2 Results and Discussion

The mean percent correct scores (hits and correct rejections) are plotted in Figure 3. A one-way ANOVA was conducted on the mean number of correct responses across the learning conditions. We found a main effect of learning condition [$F(2,46)=5.847$, $p<0.01$]. A post-hoc Newman-Keuls analysis revealed no difference between performance in the visual or haptic learning conditions alone. However, performance in the bimodal learning condition was significantly better than performance in either the visual learning [$p<0.05$] or haptic learning [$p<0.005$] conditions alone. A one-way analysis of the hit responses only revealed a similar main effect of learning condition [$F(2,46)=4.462$, $p<0.02$]. Again, a Newman-Keuls analysis revealed a significant difference between performance to the bimodal learning condition and the visual learning ($p<0.05$] and haptic learning [$P<0.02$] alone, with no differences between visual and haptic learning conditions.

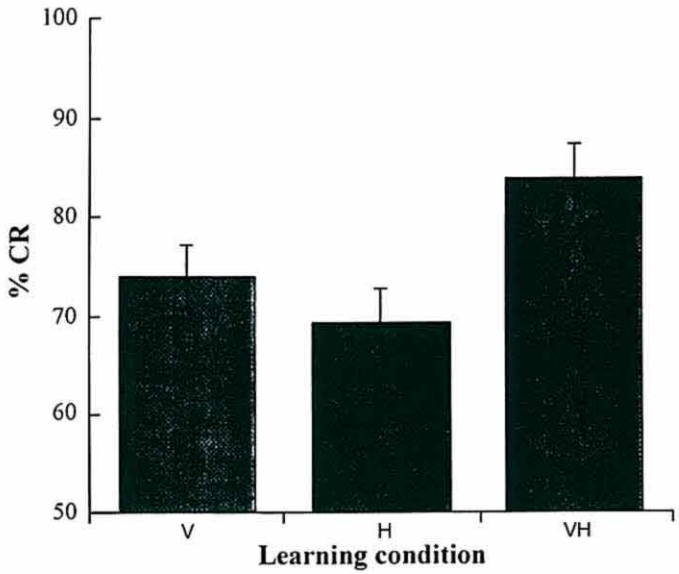

Fig. 3. Plot showing mean percent correct scores across the learning conditions in Experiment 3. Recognition was always tested bimodally (VH). Recognition performance was better when objects were learned bimodally than when learned either visually or haptically.

In this experiment we tested whether encoding properties of an object using both modalities simultaneously promoted better recognition performance than when objects were learned by each modality separately. We found that indeed, two modalities are better than one for object recognition.

5 Overall conclusions

In our study we found that recognising objects which were actively explored across modalities is less efficient than recognising these objects within the same modality. We initially thought that the cost in cross modal recognition performance was due to the poor encoding of object properties by the haptic system under limited timing. Hence we increased the time for haptic learning from 30 seconds in Experiment 1 to 60 seconds in Experiment 2. Although within-haptic performance was now equivalent to visual performance, cross-modal recognition was still less efficient than

within modal recognition. In Experiment 3 participants learned the objects using vision and touch simultaneously. Here we found the reverse effect: recognition was better in the bimodal condition than in either of the visual or haptic only conditions.

What is not clear yet from our study is how bimodal learning promotes better recognition performance. Several factors may account for this finding. First, recognition may be better simply because more information was available in the bimodal condition than in the uni-modal learning conditions. A second, and more interesting, suggestion is that each uni-modal representation is enhanced by bimodal information. In other words, the visual representation of an object may be enhanced by the encoded haptic information about the object. A third possibility is that an 'amodal' representation is richer than a uni-modal representation. One way in which to test these latter two possibilities is by testing uni-modal recognition after bimodal learning. If cross-modal learning enhances each modality then recognition performance should be the same across visual, haptic and bimodal recognition conditions. If, on the other hand, an amodal representation is created, then recognition performance should be best in the bimodal recognition condition, relative to the visual or haptic recognition conditions only. This experiment is currently in progress and the results will be discussed at a later stage.

References

1. Bülthoff, H. and Edelman, S. (1992). Psychophysical Support for a Two-dimensional View Interpolation Theory of Object Recognition. Proceedings of the National Academy of Sciences U.S.A. 89 , 60-64.
2. Edelman, S. and Bülthoff, H.H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research, 32, 2385-2400.
3. Humphrey, K. and Khan, S.C. (1992) Recognising novel views of three-dimensional objects. Canadian Journal of Psychology, 46, 2, 170-190.
4. Lawson, R., Humphreys, G.W. and Watson, D.G. (1994) Object recognition under sequential viewing conditions: Evidence for viewpoint-specific recognition procedures. Perception, 23, 595-614.
5. Newell F.N., Ernst M.O., Tjan B.S., and Bulthoff H.H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12 (1): 37-42.
6. Newell, F.N. and Findlay, J.M. (1997) The Effect of Depth Rotation on Object Identification. *Perception*, 26, 1231-1257.
7. Tarr, M.J. and Pinker, S. (1989). Mental Rotation and Orientation Dependence in Shape Recognition. Cognitive Psychology, 21, 233-282.