Multisensory Recognition of Actively Explored Objects

Marc O. Ernst and Christoph Lange
Max Planck Institute for Biological Cybernetics
Fiona N. Newell, Trinity College

Abstract Shape recognition can be achieved through vision or touch, raising the issue of how this information is shared across modalities. Here we provide a short review of previous findings on cross-modal object recognition and we provide new empirical data on multisensory recognition of actively explored objects. It was previously shown that, similar to vision, haptic recognition of objects fixed in space is orientation specific and that cross-modal object recognition performance was relatively efficient when these views of the objects were matched across the sensory modalities (Newell, Ernst, Tjan, & Bülthoff, 2001). For actively explored (i.e., spatially unconstrained) objects, we now found a cost in cross-modal relative to within-modal recognition performance. At first, this may seem to be in contrast to findings by Newell et al. (2001). However, a detailed video analysis of the visual and haptic exploration behaviour during learning and recognition revealed that one view of the objects was predominantly explored relative to all others. Thus, active visual and haptic exploration is not balanced across object views. The cost in recognition performance across modalities for actively explored objects could be attributed to the fact that the predominantly learned object view was not appropriately matched between learning and recognition test in the cross-modal conditions. Thus, it seems that participants naturally adopt an exploration strategy during visual and haptic object learning that involves constraining the orientation of the objects. Although this strategy ensures good within-modal performance, it is not optimal for achieving the best recognition performance across modalities.

Résumé La reconnaissance des formes peut se faire aussi bien par la vision que par le toucher, ce qui soulève la question sur la manière de partager l'information entre les modalités. Dans notre article, nous présentons un bref aperçu des résultats antérieurs sur la reconnaissance transmodale des objets ainsi que de nouvelles données empiriques sur la reconnaissance multisensorielle d'objets explorés de manière active. Il a déjà été démontré que, comme dans le cas de la vision, la reconnaissance haptique d'objets fixes dans l'espace est propre à l'orientation et que la performance de la reconnaissance transmodale des objets est relativement efficace lorsque les vues des objets sont assorties parmi les modalités sensorielles (Newell, Ernst, Tjan & Bülthoff, 2001). Dans le cas des objets activement explorés, à savoir sans contraintes dans l'espace, nous avons découvert une baisse dans les performances de reconnaissance transmodale relativement aux performances de la reconnaissance modale interne. À première vue, il peut sembler que cela contredise les découvertes de Newell et al. (2001). Toutefois, une analyse vidéo détaillée du comportement exploratoire visuel et haptique durant l'apprentissage et la reconnaissance a révélé qu'une vue des objets était avant tout explorée relativement à toutes les autres. Par conséquent, l'exploration visuelle et haptique n'est pas équilibrée parmi les vues des objets. La baisse de performance parmi les modalités dans le cas des objets explorés de manière active pourrait être attribuable au fait que la vue de l'objet le plus étudié n'était pas assortie adéquatement entre les tests d'apprentissage et de reconnaissance dans les conditions transmodales. Il semble donc que les sujets adoptent naturellement une stratégie d'exploration durant l'apprentissage visuel et haptique qui implique de limiter l'orientation de l'objet à une seule perspective. Même si cette stratégie donne de bons résultats en condition intramodale, elle ne permet pas d'obtenir le meilleur niveau de reconnaissance en condition intermodale.

Object recognition is often thought of as a purely visual task. However, this is by no means true. All the other sensory modalities, such as touch and audition, also contribute to the recognition of objects. Moreover, even if shape recognition is done predominantly in the visual modality, we often use our hands to actively re-orient the object with respect to the eye for recognition. Thus, everyday visual object recognition is almost never a passive task but involves active exploration. In contrast, most object recognition tasks conducted in the laboratory involve passive viewing conditions (e.g., Tarr & Bülthoff, 1999). Thus, the most natural way of exploring objects for recognition is not properly reflected in the literature. Here we aim to shed light on visu-
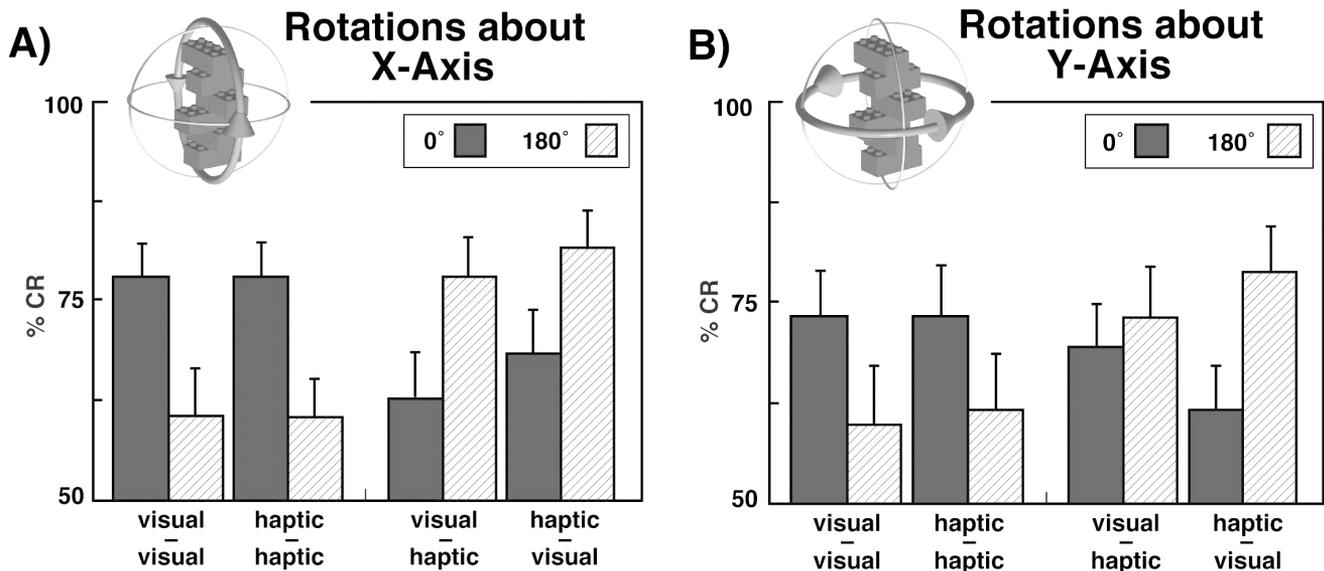
Figure 1. Percent correct recognition performance for visual and haptic shape recognition of Lego™ objects with fixed orientation. The object was either placed in the same orientation during learning and recognition test (0°) or way rotated 180° around the x-axis (A) or the y-axis (B). (Replotted from Newell et al. 2001.)

al-haptic recognition of actively explored objects.

Our world is made up of a myriad of shapes and the goal of any recognition system is to achieve what is known as object constancy: Incidental changes to an object's image characteristics, such as changes in illumination, viewpoint, or context should not affect the recognition of an object. Ever since Marr's seminal work on object recognition (Marr, 1982), there has been a flurry of activity amongst vision scientists to explain how objects are represented in visual memory such that object constancy is achieved (e.g., Biederman, 1987; Biederman & Gerhardstein, 1993; Bülthoff & Edelman, 1992; Lawson & Humphreys, 1996; Newell & Findlay, 1997; Tarr & Bülthoff, 1998). So far, however, most studies on object recognition have ignored the contribution from other sensory modalities to the representation of the object in memory. Yet information from other modalities may help to solve the object constancy problem and result in more robust recognition. In other words, a rich representation of an object may be achieved by combining or transferring information about objects across the different sensory modalities. One aim of this article is therefore to investigate whether information about objects can be efficiently shared across the visual and haptic modalities under active exploration.

A few recent studies have shown that information from one sense can affect perception in another sensory modality, often resulting in recognition performance that is robust to incidental changes in the environment.

For example, Simons, Wang, and Roddenberry (2002) reported that vestibular, proprioceptive, and optic flow information together can be used to update the representation of an object in visual memory resulting in recognition performance that is independent of viewpoint changes. Simons et al. (2002) found that the recognition of a novel view of an unfamiliar object was better when the observer moved around the object to the new view of the object than when the new view was presented to a passive observer. Similarly, Pasqualotto, Finucane, and Newell (2005) found that observer movement can also update an object's representation in tactile memory thereby eliminating effects of changes in viewpoint that are otherwise present in a passive viewing condition. Here we ask whether active exploration of the objects leads to an omni-directional representation of object shape in memory, which can be shared across the visual and haptic modalities.

Cross-modal Object Recognition

The results of cross-modal priming studies have suggested that information about objects can be efficiently shared across the visual and tactile senses (Easton, Greene, & Srinivas, 1997; Jüttner, & Rentschler, 2002; Reales & Ballesteros, 1999; Rentschler, Jüttner, Osman, Müller, & Caelli, 2002, 2004). However, because the stimuli were named or verbally described during these studies it is not clear what information, or indeed what system, mediated cross-modal priming. Both the Easton et al. (1997) and Reales and Ballesteros (1999) studies

provided some evidence for implicit or perceptual priming; nevertheless, it was still possible that verbal or semantic memory may have mediated cross-modal priming since priming occurred even when the object percepts differed (e.g., a haptic target object primed a visual line drawing).

In an effort to examine what shape information is encoded in vision and touch, respectively, in a recent recognition memory study we used unfamiliar Lego™ objects as stimuli presented in a fixed position (Newell et al., 2001) (c.f. Figure 2). Using unfamiliar Lego™ objects, we limited the possibility of a modality encoding bias due to changes in weight, size, texture, or colour between the objects. We used an old/new recognition paradigm and recorded percent correct recognition performance for objects, which were either presented in the same orientation at recognition test as during learning or which were rotated 180° around between learning and test. We tested under visual and haptic learning and testing conditions (V-V, H-H, V-H, and H-V). Figure 1 plots the recognitions results for rotations around the x-axis (A) and the y-axis (B) (replotted from Newell et al., 2001). The results of this study suggested that different aspects of an object's shape can be encoded by the tactile and visual sensory systems. Specifically, we found better haptic recognition performance for the surface of the object facing away from than towards the observer, whereas visual recognition was best for the object surface facing the observer. Moreover, when the encoded information was matched across modalities (i.e., if the back of the felt object was rotated by 180° and presented to the visual system) then no cost was observed on recognition performance, suggesting that shape information can be easily shared across the systems.

Our findings made intuitive sense since the back of the Lego™ objects is the surface that the hands naturally and predominantly explore when the objects are fixed in space with respect to the person exploring it. Furthermore, the back of an object is generally the surface that is not visible; therefore, by encoding this surface via the haptic sense, it provides complementary information about the shape of the object. If the optimal views to each modality were matched then no cost in cross-modal performance was observed (to see this compare recognition performance for 0° in the within-modal conditions with 180° in the cross-modal conditions in Figure 1). For objects that are rotationally or bilaterally symmetrical, the same information about an object may be stored in visual and tactile memory (Ballesteros, Magna, & Reales, 1997), allowing for efficient transfer of information and also shape priming across the modalities.

Active Visual and Haptic Exploration

In Newell et al. (2001), the position of the object stimulus and its orientation was always fixed in space. It is possible that when objects are actively explored (without constraining the spatial position and orientation of the object), all surfaces are encoded by both sensory systems providing omni-directional information and so causing recognition to be independent of view. We may therefore predict that spatially unconstrained, active perception would result in a rich, surface-independent representation of the objects in each modality. Consequently, we should find no cost on recognition performance when crossing modalities at test.

On the other hand, it is known that even under active perception, vision encodes certain views of objects preferentially over others to maximize recognition performance. This is often referred to as the *canonical view* of objects (Blanz, Tarr, & Bülthoff, 1999; James, Humphrey, & Goodale, 2001; James et al., 2002; Perrett & Harries, 1988). Indeed, we have recently found evidence for canonical views in haptics for both familiar and unfamiliar objects and that, like vision, these views promote more efficient haptic object recognition (Woods, Moore, & Newell, 2007). If, however, the visual and haptic systems show different preferences for representing the objects (i.e., vision and touch may, for example, represent objects in different canonical views), then the system may be faced with a problem when recognition occurs in the modality other than the encoding modality. Consequently, such an orientation-specific representation, even under active exploration conditions, may lead to a cost in recognition performance of objects learned in one modality but recognized in the other.

To discriminate between these two hypotheses, we investigated object recognition performance under spatially unconstrained visual and haptic exploration conditions. All target objects in our study were actively viewed during visual exploration or freely manipulated during haptic exploration. In the first experiment (Experiment 1a and b), we tested whether or not cross-modal recognition of objects was as efficient as within-modal recognition under active exploration of the objects. Furthermore, we directly analyzed the pattern of exploration behaviour of participants exploring the Lego™ objects either visually or haptically using video analysis.

The experimental procedures of this experiment and the stimuli used were essentially the same as used by Newell et al. (2001), with the only difference that the objects were not spatially constrained during exploration (cf. Figure 2). Thus, we used an old/new recognition task assessing object recognition performance (see Appendix for details). The objects could be
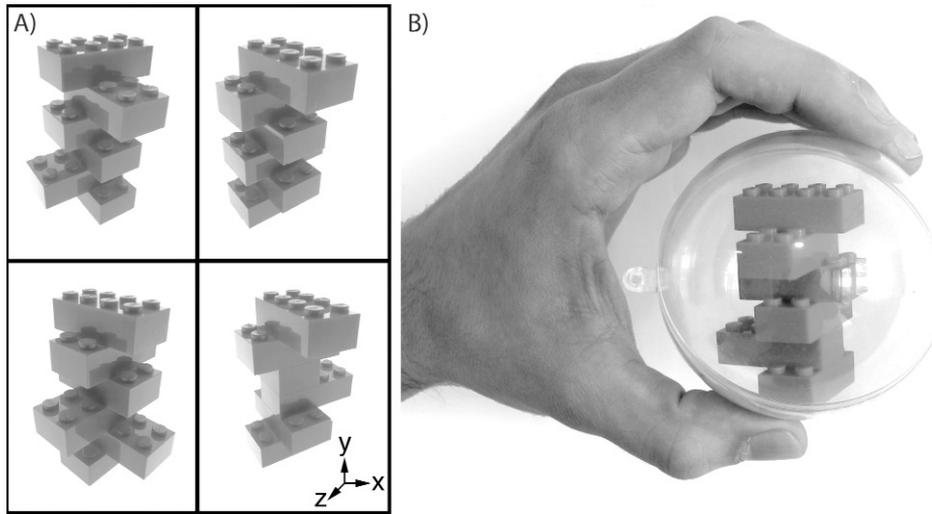
Figure 2. A) A set of example stimuli used in our experiment (replotted from Newell et al., 2001). B) One object placed in the Perspex sphere for active visual exploration.
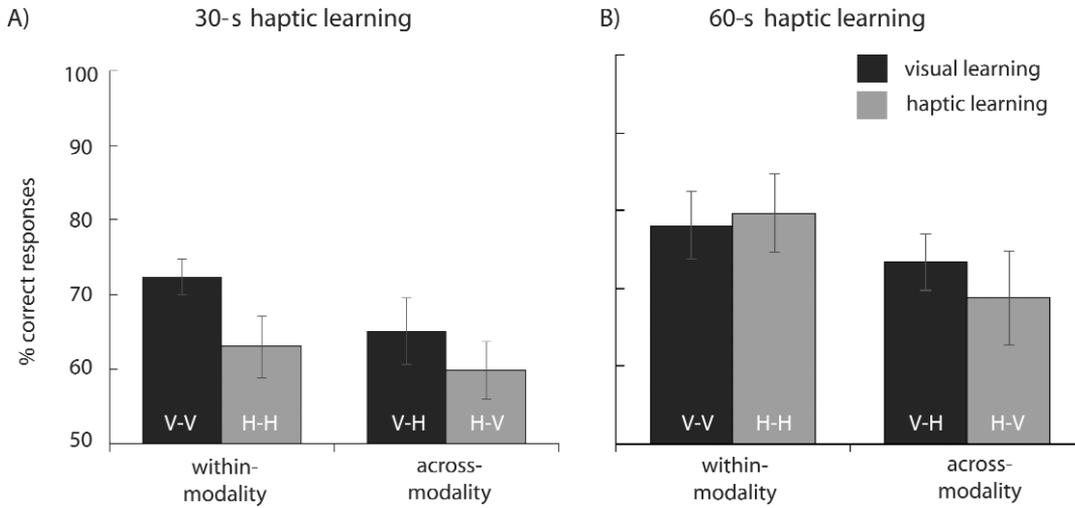

Figure 3. Plot showing mean percentage correct responses (%CR) for the within-modal and cross-modal recognition conditions. A) Results of Experiment 1a for which visual and haptic learning times were 30 s. B) Results of Experiment 1b for which haptic learning time was increased to 60 s, whereas visual learning time was kept at 30 s.

learned either visually-alone or haptically-alone and recognition performance was then subsequently tested either visually-alone or haptically-alone, resulting in two within-modal (V-V, H-H) and two cross-modal conditions (V-H, H-V). To allow for active visual exploration, we presented the Lego™ objects in a Perspex sphere (see Figure 2). This way, participants could freely view the object from all directions without touching it directly. Thus, touch could be used to re-orient

objects for visual learning or recognition but did not provide shape information (Klatzky, Lederman, & Matula, 1993).

Figure 3A shows the recognition results (% correct recognition performance) for the four testing conditions: V-V, H-H (within-modal), and V-H, H-V (cross-modal). Even though it seems there is a trend in this experiment, we did not find a significant cost in the recognition of objects learned in a different sensory

modality than the test modality (see Appendix for details on statistical tests). Given the statistical criterion, cross-modal recognition was as good as within-modal recognition. However, our failure to find evidence for a significant cost in recognition in this experiment should be treated with caution since we found an effect of sensory modality at learning: Learning an object through vision promoted better recognition performance than learning an object through touch. Furthermore, visual learning promoted better within- and cross-modal recognition performance than haptic learning, since there was no interaction between the factors.

To get a better estimate of whether or not there is a cost involved in recognition across modalities, it is important to ensure that the same amount of information is encoded in both modalities. For the encoding of shape information, haptics is generally slower than vision (e.g., Jones, 1981; Newell et al., 2001; Woods, O'Modhrain, & Newell, 2004) and consequently performance was worse for haptic than for visual learning. In our first Experiment (1a), the learning times for both vision and touch were limited to 30 s each. To ensure equivalent within-modal recognition performance, we decided to repeat this experiment but we increased the learning time for haptics-only to 60 s, leaving the learning time for vision-only at 30 s (Experiment 1b).

As can be seen in Figure 3B, the lack of interaction between learning modality and recognition condition indicates that our learning time manipulation had the desired effect. By increasing the time to learn an object using touch relative to vision, we succeeded in rendering recognition performance equal across both within modality conditions. With this manipulation, however, we now found a significant cost in cross-modal recognition relative to within-modal recognition. Performance dropped significantly from on average 78.9% correct in the within-modality condition to 71.1% in the cross-modal condition (see Appendix for details). If performance across the within-modal conditions was equivalent, we might ask what caused the cost in crossing modalities at test? As mentioned in the introduction, in the study reported by Newell et al. (2001), we did not find such a cost of transfer when the stimuli were fixed in space and the optimal "views" were matched across modalities (comparable to Experiment 1b in Newell et al. (2001), learning time was adjusted to reach equal within-modal performance for vision and touch). Therefore, we are left with the question of where this apparent discrepancy in findings comes from? To find an answer why exploration of an unconstrained object leads to a cost in cross-modal recognition performance relative to when the objects were fixed in space, in the next experiment we investigated participants' visual and haptic exploration strategies during learning and recognition test. This was achieved by videotaping the participants' pattern of exploratory behaviour while they performed the task and analyzing which "views" of the objects were preferentially explored.

Exploration Strategies

In order to study the exploration strategies observers adopted while actively exploring the objects visually or haptically, we videotaped the participants while they performed the old/new recognition task (see Appendix for details). An analysis of these videotapes revealed that when exploring the objects either visually or haptically, all participants held the Lego™ objects predominantly "upright," that is, with the top surface of the Lego™ bricks pointed upwards and the elongated axis of the object perpendicular to the ground. This natural exploration strategy constrained the sides to be explored to the four side views of the object, which largely simplified the video analysis.

Subsequently, during the video analysis, we labelled the four side views of the objects as follows: The view that was predominantly explored during learning was called the 0° view; the other three sides were called 90°, 180°, and 270° views (in clockwise direction from the 0° view). For the recognition test, we kept the labelling of the sides of the target objects from the learning phase. That is, the 0° view of the object was the same side during learning and test. The same was true for the other three views. This labelling of the views was done for all target objects individually for each participant. This way we could compare whether the side that was explored most during learning was also explored most during recognition test.

Interestingly, even though participants were free to explore the object from all sides, not all four side-views were explored with equal probability. In contrast, it seems that participants arbitrarily chose one side of the object to study predominantly during the learning phase (roughly 70% of the total exploration time) and they almost completely ignored the other three sides. This can be seen from the learning conditions plotted in Figure 4 (the front row of bars in all four panels). The same pattern of exploration strategies was found during learning whether the objects were explored visually or haptically (left and right columns of Figure 4).

Furthermore, although participants knew in advance whether they would be tested in the within- or cross-modal condition, the pattern of results in the learning condition was the same irrespective of these instructions (top and bottom rows of Figure 4). This suggests that participants did not adjust their learning strategies
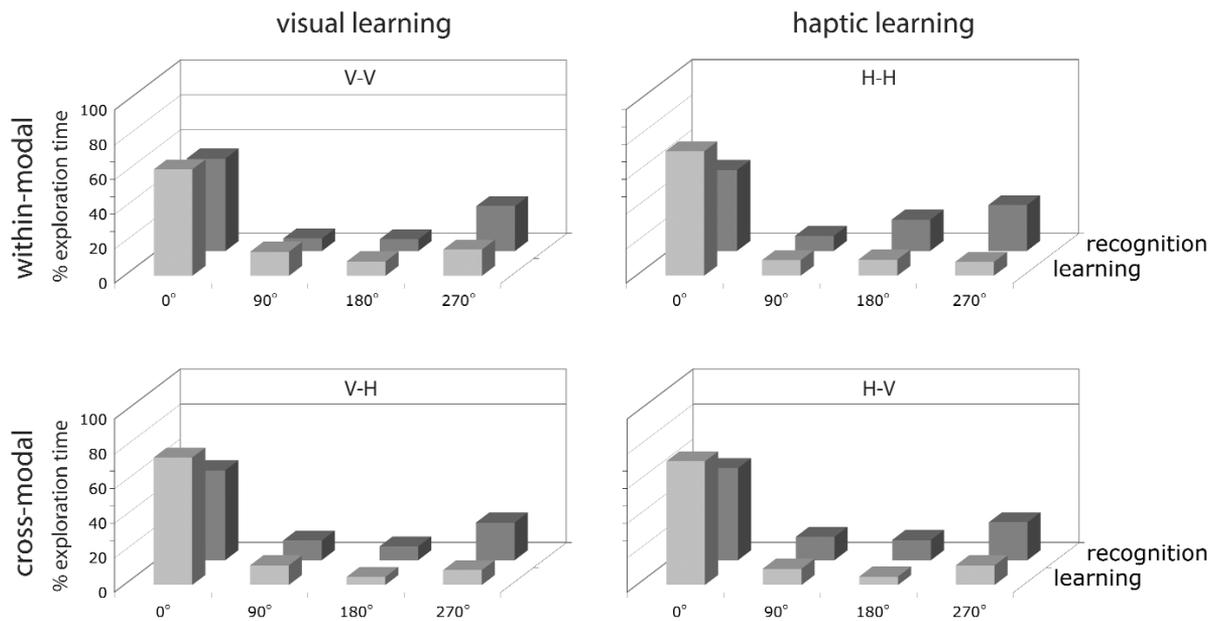
Figure 4. Percent exploration time derived from a video analysis for the four different sides during learning and recognition test. Shown are the two within- and cross-modal conditions for visual and haptic learning. The different sides (0°, 90°, 180°, and 270°) refer to the orientation of the object relative to the observer exploring it. The side that was explored most during learning was defined as the 0° orientation. The other sides were labelled accordingly in clockwise direction. The 0°, 90°, 180°, and 270° sides of each object are the same for learning and recognition.

according to the recognition conditions. That is, neither learning modality nor knowledge about the recognition test had a significant influence on participants' exploration strategies during the learning phase (see Appendix for details).

In some respects, even more interesting is the pattern of results observed in the recognition conditions. In the within-modal conditions both for vision and touch, the "0°" side of the target objects was again explored most. With roughly 50% of the relative exploration time, this is again the dominant side. The other three sides of the target objects were explored significantly less often during the recognition test (nondominant sides: 10% for "90°," 15% for "180°," and 25% for "270°" – see Appendix for details). However, these values for the nondominant sides are slightly higher than during learning. There may be several reasons for this slight increase in the relative exploration times of the nondominant sides in the recognition phase. One reason may be that it takes participants longer to search for the dominant (learned) side during recognition than to pick a side to be learned during the learning phase. Secondly, during the recognition test, the exploration time was unlimited and participants could answer as soon as they could classify the object as new or old. In other words, they could stop exploring the object as soon as they were certain about their decision. This

may render the exploration times for the dominant side (i.e., the learned side of the target objects) to be proportionally less and consequently emphasizes exploration time on the nondominant sides. Furthermore, the freely timed exploration time during recognition was on average only 15.2 s (compared with 1 min. for learning). Thus, this also may distort the proportion of time each side is explored during recognition.

Most importantly, the same pattern of results in the recognition condition is found whether the test occurred within modalities or across the modalities. This is interesting because, in order to get the best performance possible, participants should have rotated the object by 180° so that the object is turned back-to-front (see the results of Newell et al., 2001 for comparison). This re-orienting of the object in the cross-modal conditions, however, did not occur. Instead, participants tended to explore the objects in the same orientation during recognition as the orientation explored in the learning phase, independent of whether the test occurred within or across the modalities.

Taken together, this exploration behaviour explains the drop in performance for cross-modal recognition as compared to within-modal recognition. That is, even when free to explore the object from all sides, participants naturally choose a highly view-dependent exploration strategy both during learning and in the recogni-

tion test, at least with the objects used here. This, together with the fact that participants do not rotate the objects back-to-front between learning and recognition test in the cross-modal condition but always explore them predominantly from the same orientation, may contribute to the reduction in cross-modal recognition performance (as we previously argued in Newell et al., 2001).

Discussion

The purpose of this article was to shed light onto multisensory (i.e., visual-haptic) shape recognition for actively explored objects that are not fixed in space. Using an old/new recognition paradigm, we found evidence that cross-modal recognition was less efficient relative to within-modal recognition of unfamiliar objects under spatially unconstrained exploration conditions. Using a video analysis, we showed that the pattern of exploration behaviour in vision as well as in touch are typically orientation specific, in that not all sides of an object are equally explored, even when participants are free to do so. Furthermore, by using a video analysis, we could attribute the cost in cross-modal recognition performance to participants' inability to rotate the object back-to-front between learning and recognition test, which according to Newell et al. (2001) would have been necessary to achieve recognition performance that did not incur a cost. Taken together, the experimental results reported here suggest that even though there is a cost in cross-modal recognition performance, this should not be taken as evidence that shape information from vision and touch is represented differently and thereby causing the cost. Rather, it is the inability of the participants to orient the object in a way to best extract the information necessary for recognition without cost.

In general, participants oriented the target objects much in the same way during learning and recognition test (see Figure 4). Thus, certain information must be available about how to orient the object without necessarily affecting the recognition of its shape. One way this could have occurred is by using an orientation strategy that is independent of the particular object's shape. Developing such a strategy is relatively easy with the Lego™ objects used for this experiment. Consider the fact that prior to detailed exploration of the object shape, all participants consistently oriented the object such that the bumps of the Lego™ bricks pointed upwards – this constrained exploration to mainly the four sides of the object. If participants then, in addition to orienting the objects upright, used another strategy to orient the objects – for example, orient the long side of the lower brick parallel (or perpendicular) to their body axis – then this effectively would

constrain the orientation of all objects, so that there are only two possible orientations left for exploration. With just one other simple strategic rule to orient the objects, the orientation of all objects would be uniquely defined without knowing their exact shapes, which would be necessary for correct recognition. Such strategies are possible here with the Lego™ objects because they all have a similar structure. In sum, we cannot exclude the possibility that participants here used such an orienting strategy before encoding specific shape details of the object.

However, there may also be other strategies to orient an object without encoding details of the individual shape, which could be applied more generally to object recognition. For example, it is known that for many object classes there are canonical views (Blanz et al., 1999; James et al., 2001; James et al., 2002; Perrett and Harries, 1988). This is true even for unfamiliar object classes such as machined tool parts (Perrett, Harries, & Looker, 1992). Furthermore, an object can be oriented into a particular canonical view without knowing its identity based on the particular object's shape information. For example, one can orient toy cars in a canonical orientation without identifying the particular kind of car based on its unique shape. We have recently found that such canonical views (orientations) also exist for the unfamiliar objects like the Lego™, objects used for this experiment (e.g., Woods et al., 2007) but what is not known is whether such canonical views are dissimilar across the modalities that would then contribute to a cost in cross-modal recognition performance. More research is needed to answer these questions.

Although our study was not designed to address the issue of what mechanism mediates cross-modal recognition, it is interesting to speculate on how object information might be shared or combined across vision and touch. Three candidate models of how information can be shared across the senses are generally proposed (Meredith, 2002; Stein & Meredith, 1993). First, information could be held independently in modality-specific formats but accessible by another amodal system. Second, information may be recoded directly from one modality into another. For example, objects encoded through touch may be recoded into visual images. Finally, all information may be held in some amodal or multisensory form. We cannot distinguish between any of these possibilities. However, we can say from this study together with the study by Newell et al. (2001) that with the particular Lego™ objects used and when views are matched there is no cost in transferring information from one modality to the other. Although this finding does not inform us about the specific format in which the object information is represented by the dif-

ferent modalities, it does tell us that the information can be used very efficiently.

Much recent evidence from neuroimaging has supported the idea that tactile object information is recoded into a visual code (Sathian & Zangaladze, 2001; Zhang, Weisser, Stilla, Prather, & Sathian, 2004) because tactile processing of objects generally activates visual object recognition areas (although see Reed, Shoham, & Halgren, 2004 for evidence of an independent cortical route for tactile object recognition). Furthermore, it is proposed that visual imagery may mediate this recoding of tactile information (Sathian & Zangaladze, 2002). However, these studies make the assumption that areas active during object recognition are nominally "visual." Yet there is a growing body of evidence to suggest that areas once considered uni-modal, even primary sensory areas, actually respond to stimulation from other senses (e.g., Wallace, Ramachandran, & Stein, 2004). These findings challenge the notion of uni-sensory cortical areas per se and suggest that information may in fact be held in a multisensory code that is readily accessible to different modalities. Other psychophysical studies have suggested that information from across the sensory systems is integrated into this multisensory code in a statistically optimal way (e.g., Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Helbig & Ernst, 2007). Considering the results of these previous studies together with the results about the efficient use of information across the modalities for object recognition found here, we may speculate that there is a uniform multisensory code for shape perception.

In this article, we highlighted the way multisensory information is shared across the modalities and how it can be transferred from one modality to the other. Recently, there was quite some research activity on the question of how multisensory information is integrated across the senses. It was shown that for simple object properties such as, for example, visual and haptic size (Ernst & Banks, 2002) or visual and auditory location (Alais & Burr, 2004), the integration of multisensory information is statistically optimal. Statistically optimal here means that the combined multisensory estimate has the smallest variance possible and is thus the most reliable estimate that can be achieved given the information provided by the sensory modalities. Also for slightly more complex objects, such as ellipses, the integration process has already been demonstrated to be optimal (Helbig & Ernst, 2007). However, the question of integration is slightly different from the question addressed here in this article. For integration to occur, two redundant sources of information have to be available (Ernst & Bülthoff, 2004), which are then combined into one unique multisensory representation of the object property. In contrast, here we investigated how information from one modality can be used for recognition in another modality. It would be interesting, though, to see how multisensory recognition performance changes when both modalities are available at the same time so that integration of multisensory information could occur. The optimal integration hypothesis would predict that recognition performance should increase when both modalities (vision and touch) are available simultaneously. However, with more complex objects, such as the Lego™ objects used here and many familiar objects in general, the predictions are not so straightforward. First, the gathering of haptic information through manual exploration is rather sequential and for vision it is rather parallel. Second, while exploring the objects, the hand might occlude the object from sight. All these are issues that have to be considered when investigating optimal multisensory integration for object recognition. How all these factors affect the integration process of complex multisensory object information is still unclear and remains an open field of research.

In sum, our experimental results show that visual-haptic object perception is orientation specific even when participants are free to actively explore the objects from all directions. Interestingly, such an orientation-specific exploration strategy is adopted even when it comes at a cost as demonstrated here for cross-modal shape recognition. Our findings suggest that under natural or unconstrained conditions, the nature of the shape information encoded through either vision or touch is not random but is a subset of possible shape information limited to a single snapshot or view.

Please address correspondence to Marc O. Ernst,, Max Planck Institute for Biological Cybernetics, Spemannstr. 41, Tübingen, Germany (Tel: +49 7071 601 644; Fax: +49 7071 601 616; E-mail: marc.ernst@tuebingen.mpg.de).

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257-262.

Ballesteros S., Manga D., & Reales J. M. (1997). Haptic discrimination of bilateral symmetry in 2-dimensional and 3-dimensional unfamiliar displays. *Perception & Psychophysics*, *59*, 37-50.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review 94*, 115-147

Biederman I., & Gerhardstein (1993). Recognizing depth-rotated objects: Evidence and conditions for three-

dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1162-1182.

Blanz V., Tarr M. J., & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception, 28*, 575-599.

Bülthoff, H. H., & 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, I., & Newell, F. N. (2004). Interactions between vision and audition for face recognition. *Perception, 33*, 108.

Easton, R. D., Greene, A. J., & Srinivas, K. (1997). Transfer between vision and haptics: Memory for 2-D patterns and 3-D objects. *Psychonomic Bulletin and Review, 4*(3), 403-410.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429-433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162-169.

Helbig, H. B., & Ernst, M. O. (2007) Optimal integration of shape information from vision and touch. *Experimental Brain Research 179*, 595-606.

James, K. H., Humphrey, G. K., & Goodale, M. A. (2001). Manipulating and recognizing virtual objects: Where the action is. *Canadian Journal of Experimental Psychology, 55*, 111-120.

James, K. H., Humphrey, G. K., Vilis, T., Corrie, B., Baddour, R., & Goodale, M. A. (2002). "Active" and "passive" learning of three-dimensional object structure within an immersive virtual reality environment. *Behaviour Research Methods, Instruments and Computers, 34*, 383-390.

James, T. W., Humphrey, G. K., Gati, J. S., Servos P., Menon, R. S., & Goodale, M. A. (2001). Haptic study of three dimensional objects activates extrastriate visual areas. *Neuropsychologia, 40*, 1706-1714.

Jones, B. (1981). *The developmental significance of cross-modal matching.* In R. Walk & H. Pick, Jr. (Eds.), Intersensory perception and sensory integration (pp. 109-136). New York: Plenum.

Jüttner, M., & Rentschler, I. (2002). Imagery in multi-modal object learning. *Behavioral and Brain Sciences 25*, 197-198.

Klatzky, R. L., Lederman, S. J., & Matula, D. E. (1993). Haptic exploration in the presence of vision. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 726-743.

Lawson, R., & Humphreys, G. W. (1996). View specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 395-416.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: W. H. Freeman.

Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research, 14*, 31-40.

Newell, F. N., Ernst, M. O., Tjan, B. J., & Bülthoff, H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science, 12*, 37-42.

Newell, F. N., & Findlay, J. M. (1997). Effects of depth rotation on object identification. *Perception, 26*, 1231-1257.

Pasqualotto, A., Finucane, C., & Newell, F. N. (2005). Visual and haptic scenes are updated with observer motion. *Experimental Brain Research. 166*(3-4), 481-488.

Perrett, D. I., & Harries, M. H. (1988). Characteristic views and the visual inspection of simple faceted and smooth objects: Tetrahedra and potatoes. *Perception, 17*, 703-720.

Perrett, D. I., Harries, M. H., & Looker, S. (1988). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception, 21*, 497-515.

Reales, J. M., & Ballesteros, S. (1999). Implicit and explicit memory for visual and haptic objects: Cross-modal priming depends on structural descriptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 644 -663.

Reed, C. L., Halgren, E., & Shoham, S. (2004). The neural substrates of tactile object recognition: An fMRI study. *Human Brain Mapping, 21*, 236-246.

Rentschler, I., Jüttner, M., Osman, E., Müller, A., & Caelli T. (2002). Multimodal representations for human 3D object recognition. In R. P. Würtz & M. Lappe (Eds.), *Dynamic perception* (pp. 327-332). Amsterdam: IOS Press.

Rentschler, I., Jüttner, M., Osman, E., Müller, A., & Caelli, T. (2004) Development of configural 3D object recognition. *Behavioural Brain Research, 149*, 107-111.

Sathian K., & Zangaladze A. (2001). Feeling with the mind's eye: The role of visual imagery in tactile perception. *Optometry and Vision Science, 8*, 276-281.

Sathian, K., & Zangaladze, A. (2002). Feeling in the mind's eye: Contribution of visual cortex to tactile perception. *Behavioural Brain Research, 135*(1-2), 127-132.

Simons, D. J., Wang, R. F., & Roddenberry, D. (2002). Object recognition is differentially affected by display orientation and observer viewpoint changes. *Perception & Psychophysics, 64*, 521-530.

Stein, B. E., & Meredith, M. (1993). *The Merging of the Senses.* Cambridge, MA: MIT Press.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67*(1-2), 1-20.

Tarr, M. J., & Bülthoff, H. H. (1999). *Object recognition in*

man, monkey, and machine. Cambridge, MA: MIT Press.

Wallace M. T., Ramachandran R., & Stein B. E. (2004). A revised view of sensory cortical parcellation. *Procedings of the National Academy of Science, 101*(7), 2167-72.

Woods, A. T., Moore, A., & Newell, F. N. (2007). *Canonical views in haptic object recognition*. Manuscript submitted for publication.

Woods, A. T., O'Modhrain, S., & Newell, F. N. (2004). The effect of temporal delay and spatial differences on cross-modal object recognition. *Cognitive, Affective & Behavioral Neuroscience, 4,* 260-269.

Zhang, M., Weisser, V. D., Stilla R., Prather, S. C., & Sathian, K. (2004). Multisensory-cortical processing of object shape and its relation to mental imagery. *Cognitive, Affective & Behavioural Neuroscience 4,* 251-259.

Appendix

Experiment 1: Active Visual and Haptic Exploration

Participants

For Experiment 1a (Figure 3a), 24 individuals (7 female/17 male) were recruited from the Max Planck Institute for Biological Cybernetics subject pool to participate in the experiment for pay (8 Euro per hour). Their ages ranged from 27 to 39 years. In Experiment 1b (Figure 3b), 18 participants (5 female/13 male) with range in age between 22 and 36 years took part. All participants were naïve to the purposes of the task and all reported normal or corrected-to-normal vision. None reported any tactile impairment. All participants gave written consent prior to conducting the experiment.

Stimuli

We used the same stimulus set of unfamiliar objects as described in Newell et al. (2001): All objects were composed of six, same-sized Lego bricks (the 8-bit brick measuring 3 cm x 1.5 cm each) stacked on top of each other, so that long and short faces were alternating. Each object was defined by a unique configuration of these bricks and had a middle vertical column of 4 bits (see Fig. 2 left panel for an illustration of some example objects). We choose the configuration of the bricks in a way that there was symmetry along the x-, y-, or z-axes. All of the bricks were red in colour. We created 32 individual object stimuli for the experimental sessions and 4 extra objects for the practice session.

For the visual conditions, each object was placed inside a transparent Perspex sphere (Fig. 2, right panel). The diameter of a sphere was 10 cm into which a single object neatly fitted. For the haptic conditions, each object stimulus was placed directly onto the participant's hands behind a curtain thus obscuring the haptic targets from view.

Design

To test recognition performance, we used an old/new recognition paradigm. The experiment was based on a two-factor repeated measures design with learning sensory modality (vision or haptics) and recognition conditions (within- or cross-modality) as factors. The experiment was divided into four separate blocks with different learning and testing conditions, two within-modal: visual-visual (V-V) and haptic-haptic (H-H), and two cross-modal conditions: visual-haptic (V-H) and haptic-visual (H-V). The order of these blocks was counterbalanced 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 nontarget (distractor) within a block.

Procedure

Prior to each experimental block, participants were informed of the sensory modality they would use to learn the objects and the sensory modality, which would be tested. Each of the two experiments was divided into four separate experimental blocks and participants could take a self-timed break between each block. The task for the participant was an old/new recognition task. Participants had to learn four target objects presented one at a time in random order. During recognition test, we sequentially presented the target objects amongst four distractor objects and participants had to classify each object as old or new. During the first eight trials of the recognition test, we presented the four target objects and the four distractor objects in random order. Additionally, at the end of each block, we presented four extra repeats. Thus, in each test block there were 12 trials. The four extra repeats were presented last and could either be target or distractor objects. We repeated four trials in order to avoid our participants using a guessing strategy by, for example, counting the number of times they said "old." Participants were instructed that all four learned objects would be presented at least once (or multiple times) during test and there was no instruction about the order of presentation. Data from these repeated trials were not included in the data analyses. The experiment was preceded by four practice trials, one from each block. For Experiment 1a, participants were given 30 s to learn each object during learning, irrespective of modality. In Experiment 1b, we adjusted the learning times to achieve

equal recognition performance in the two within-modal conditions (V-V and H-H). This was done be giving participants 30 s to visually learn each objects and 60 s to haptically learn each object. The learning time was the only difference between Experiment 1a and b. During recognition test, participants were given unlimited time to respond after each trial. Responses were given verbally and noted by the experimenter. Each experiment took approximately one hour for each participant to complete.

Statistical Tests

The mean percent correct scores (hits and correct rejections) for both Experiments are plotted in Figure 3. A two-way ANOVA was conducted on the mean number of correct responses across the learning and recognition conditions. In Experiment 1a (Figure 3a) during which all learning times were 30 s, we found no main effect of recognition condition, $F(1, 23) = 2.19$, *n.s.* 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, $F(1, 23) < 1$, *n.s.* An analysis of the hit trials only (i.e., correct targets identified) showed the same pattern of results; no effect of recognition condition, $F(1, 23) < 1$, *n.s.*, a main effect of learning, $F(1, 23) = 6.05$, $p < 0.05$, and no interaction, $F(1, 23) = 2.19$, *n.s.*

A detailed description of the results of Experiment 1b are provided in Figure 3b. Here the mean percent correct scores (hits and correct rejections) are plotted. A two-way ANOVA was conducted on the mean number of correct responses across the learning and recognition modalities. Here there was no effect of learning modality, $F(1, 23) < 1$, *n.s.* and again no interaction between the factors, $F(1, 15) < 1$, *n.s.* However, as speculated above, we now found a significant main effect of recognition condition, $F(1, 15) = 7.65$, $p < 0.05$, indicating that performance was better for within-modal recognition than across modalities. An analysis of the hit trials only (i.e., correct targets identified) showed the same pattern of results: a main effect of recognition condition, $F(1, 15) = 10.08$, $p < 0.01$, no effect of learning, $F(1,15) < 1$, *n.s.*, and no interaction, $F(1, 15) < 1$, *n.s.*

Finally, an analysis of the cross-modal condition using a sign-test revealed that performance was significantly better than chance for all tested conditions, $Z = 3.098$, $p < 0.002$.

<div align="center">Experiment 2: Exploration Strategies</div>

Participants

We recruited six individuals to participate in Experiment 2 for pay (8 Euro per hour). Three of the participants were female and three male. Their ages ranged from 22 to 34 years. All participants were naïve to the purposes of the task, they did not participate in any of the previous tasks, and all reported normal or corrected-to-normal vision. None reported any tactile impairment. All participants gave written consent prior to conducting the experiment.

Stimuli and Procedure

We used the same set of stimuli as in the previous experiments. The behavioural task was exactly the same as in Experiment 1. That is, we conducted an old/new recognition task with the four conditions, V-V, H-H (within-modal) and V-H, H-V (cross-modal). The learning times were constrained to 60 s in both the visual and haptic learning conditions.

We used a mini DV-camera to videotape the participants during all the learning and recognition conditions. The camera was mounted opposite the participant, and was angled slightly top-down onto the participants' hands whilst they explored the objects. As previously described, in the visual learning or recognition condition, the Lego™ object was fitted into a Perspex sphere so the hands did not directly touch the objects while visually exploring it.

Analysis of Video Sequences

With the particular Lego™ objects we used, the bottom and top views were not very informative and so these views were relatively unexplored by the participants. Therefore, we constrained our video analysis to the four side views of each object. To simplify the video analysis, we put uninformative visual markers on the Lego™ bricks, which if viewed did not provide any direction or identity-specific information to the participant that could help to perform the recognition task. From the video sequences we extracted the duration of time a particular side of an object was facing the observer during exploration. For each of the four sides facing the observer during exploration, we recorded the start and end times the participant studied a particular side. If none of the four sides directly faced the observer (e.g., for top and bottom views), the object's orientation was classified as "other." Relative to the entire exploration time, the length of time classified as "other" was on average below 4% and thus was negligible. For each side, we calculated the percentage of time it was explored with respect to the other four sides. This was done both for the learning phase and the recognition test. The side of the object that was explored the longest during the learning phase was labelled "0°." The other three sides were labelled "90°," "180°," and "270°" in clockwise direction (as seen from the participants' view) with the object in upright orientation. This was done for each object and for each participant individually. The relative times for the sides labelled this way were then averaged.

Statistical Tests

During learning roughly 70% of the entire exploration time was spent on one side of the object only and roughly 10% was allocated to each of the other three sides. This shows that only one side was effectively explored, ANOVA $F(3, 15) = 35.24$, $p < 0.05$, Tukey-HSD for each comparison of side "0°" with another side $p < 0.01$ and that all other sides were treated equally (Tukey-HSD for all pairwise comparisons of sides "90°," "180°," and "270°" $p > 0.05$.

Neither in the within-modal nor in the cross-modal condition was there any difference between visual or haptic learning, $F(1, 5) = 2,48$, $p = 0.176$ and $F(1, 5) = 1,62$, $p = 0.26$, respectively. Furthermore, there was no interaction between these factors.

During recognition test the dominant side was roughly explored 50% of the time. The other three sides of the target objects were explored significantly less often (nondominant sides: 10% for "90°," 15% for "180°," and 25% for "270°": ANOVA $F(3, 15) = 5.99$, $p < 0.05$, Tukey-HSD < 0.05 for each difference between "0°" and any other side, > 0.10 for any other comparison. There was neither a significant effect of learning modality or recognition condition, nor was there any interaction between these two factors.