Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Physiology - Paris 98 (2004) 147–159

Journal of Physiology Paris

www.elsevier.com/locate/jphysparis

Visual, haptic and cross-modal recognition of objects and scenes

Andrew T. Woods, Fiona N. Newell *

Department of Psychology, Trinity College, University of Dublin, Aras an Phairsaigh, Dublin 2, Ireland

Abstract

In this article we review current literature on cross-modal recognition and present new findings from our studies on object and scene recognition. Specifically, we address the questions of what is the nature of the representation underlying each sensory system that facilitates convergence across the senses and how perception is modified by the interaction of the senses.

In the first set of our experiments, the recognition of unfamiliar objects within and across the visual and haptic modalities was investigated under conditions of changes in orientation (0° or 180°). An orientation change increased recognition errors within each modality but this effect was reduced across modalities. Our results suggest that cross-modal object representations of objects are mediated by surface-dependent representations. In a second series of experiments, we investigated how spatial information is integrated across modalities and viewpoint using scenes of familiar, 3D objects as stimuli. We found that scene recognition performance was less efficient when there was either a change in modality, or in orientation, between learning and test. Furthermore, haptic learning was selectively disrupted by a verbal interpolation task. Our findings are discussed with reference to separate spatial encoding of visual and haptic scenes.

We conclude by discussing a number of constraints under which cross-modal integration is optimal for object recognition. These constraints include the nature of the task, and the amount of spatial and temporal congruency of information across the modalities.

Keywords: Cross-modal; Vision; Haptics; Object recognition; Scene recognition

1. Introduction

Recognising an object for the first time is a complicated process, yet it is achieved with surprising accuracy. In visual processing, saccades scan across an object, analysing its colouring, size and shape. In haptics, the hand can feel the surface texture, encoding tiny bumps and grooves whilst simultaneously calculating how much pressure is needed to stop the object from falling to the floor. What is felt is combined with what is seen and a search through memory for translucency, coldness, smoothness and a hollow, cylindrical shape reveals that you are holding a glass. Both the visual and haptic systems together can provide clues to the identity of the object; both perceive a number of similar and a number of dissimilar object features, and through combining this information the object can be successfully identified. On its own, vision could have deduced that it was a glass, but would not have provided the information necessary for the prevention of drinking ice-cold water. Similarly touch could also have identified the object as a glass but it would not forewarn you that the contents are ice-cold tea rather than water. Thus, vision and haptics work together to create a rich, cross-modal representation of the glass and its contents.

This brief anecdote demonstrates the co-dependence between modalities such as vision and touch in object recognition. For the purpose of object recognition, both vision and haptics can be considered as image-based, in that both modalities can acquire shape information for recognition, albeit using different means. Depending on the nature of the goal, vision can dominate touch, touch can dominate vision, or if information is equally reliable, then they can both contribute to the percept. An emergent property of the brain is the integration of this otherwise disparate information (see Fig. 1) in order to provide a rich representation of objects in memory [44].

In this paper we review recent studies investigating behavioural and neural correlates of cross-modal object recognition. We have deliberately concentrated on the modalities of vision and haptics only, mainly because it

* Corresponding author. Tel.: +353-1-608-3914; fax: +353-1-671-2006.

E-mail address: fiona.newell@tcd.ie (F.N. Newell).

0928-4257/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2004.03.006

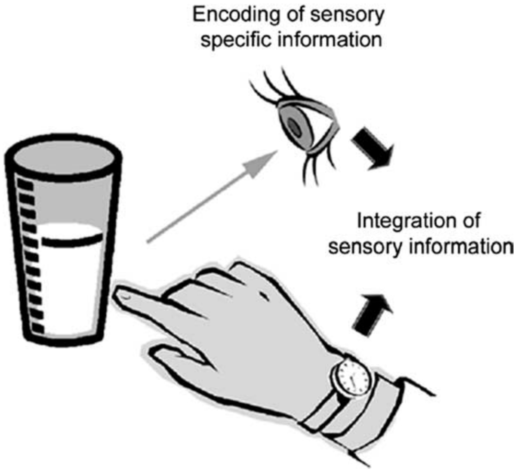
Encoding of sensory specific information
Integration of sensory information

Fig. 1. A stimulus in the environment emits a range of information that can be measured by the different sensory receptors of the central nervous system. In the diagram, light energy radiates off an object, e.g. glass, onto the retina allowing the process of perception to begin. At the same time, the surface of the skin in the fingers deforms when it comes into contact with the glass. These sensory specific codes may then be combined in the brain and to form a rich representation of an object. In our example here, a single, transparent, cold and smooth glass is recognised.

is only these modalities that can directly determine shape information for object recognition purposes (although there is evidence that audition can determine the crude shape and size of objects [18,19] such perception is indirect and is therefore not considered here). We are particularly interested in investigating how information about shape is shared across modalities, whether cross-modal recognition is efficient relative to within-modal recognition and finally how large-scale information, such as scene recognition is shared across modalities. To start with however, we briefly review recent literature on processing within these two modalities separately. We then discuss the candidate cortical areas involved in cross-modal object recognition. Finally, we describe our own studies on cross-modal object and scene perception and suggest factors affecting efficient cross-modal recognition.

2. Behavioural investigations of visual and haptic perception

2.1. Visual object recognition

An outstanding achievement of human vision is the capacity for rapid and seemingly effortless recognition of objects. Generally, the problem for the visual system is that recognition has to be achieved despite variation in the sensory information about an object. Sources of variation in an object's image on the retina, for example, can include changes in viewpoint, changes in shape with non-rigid movement, or changes in illumination. Yet the visual system must allow for such changes whilst also maintaining the uniqueness of the object's representation against other similar objects in memory (as, for example, in face recognition). The problem for vision, therefore, is that it has to achieve what is known as 'object constancy' in order to recognise an object. Visual object recognition has been the subject of much recent work by both experimental and theoretical researchers. An extensive review of the literature on visual object recognition is beyond the scope of this paper but further information can be found in [11,58,60].

Recognising configurations of multiple objects, as in scene recognition, may involve different processes than those involved in single object recognition. When viewing a scene, something of interest may attract the eye, initiating a saccade which moves the projected image of the object from peripheral vision to the fovea. Fine grain information about the object can now be accessed by the most sensitive area of the eye. Depth cues are used to deduce the object's size whereas other information such as texture and colouring give clues to material it comprises. Whilst each object is being scrutinised the larger context, i.e. the entire scene, must also remain in memory. The general contents of a scene are crudely perceivable at a glance and can be readily interpreted [3]. However, recent research into 'change blindness' has shown that the consequent representation of the scene from a glance is neither rich nor detailed (see [54] for a review). Nevertheless, one advantage of being able to perceive an entire scene rapidly is that the configuration of objects and their relative sizes are automatically available which may provide a context for subsequent recognition tasks.

2.2. Haptic object recognition

We naturally presume that touch is a slow, inadequate system for exploring the world and research on passive touch tends to confirm this notion. However, when we engage in haptic perception, that is, when we actively explore an object with our hands, information about the object is more easily accessible. Haptic perception (from the Greek word 'haptikos' meaning 'to touch') incorporates both touch information from the skin and kinaesthetic information from the position and movement of the joints and muscles. Compared to visual recognition however, our understanding of how object recognition is achieved by the haptic system is relatively poor.

Gibson [20] offers some insights into the problems the touch system has to overcome in order to achieve recognition: for example, touch also has to solve the problem of object constancy: that is, even though a single object which is being manipulated by a hand stimulates many different receptors across five fingers and a palm, it is still perceived as a single, stable entity, with a uniform surface stiffness. In order to achieve this percept the brain must integrate different information from across the five fingers and palm of the hand to form an adequate representation of the held object (for a review of the mechanics behind the sensation of touch the reader is directed to [5]). This is achieved in the brain because it is expected that the same object can stimulate receptors in adjacent fingers. The well-known Aristotle illusion nicely demonstrates this fact [43]. When an object (e.g. a pencil) is pushed between two adjacent fingers one object is perceived with two sensory impressions made by the two opposite sides of the object. The information from the haptic system therefore specifies one object. But if that same pair of fingers is crossed so that non-adjacent areas of the skin are opposed then the illusion of feeling two objects is often perceived.

Gibson argued that active touch is essential for efficient acquisition of information about an object. In one of the classical experiments on touch, Gibson [20] asked participants to identify the shape of different biscuit cutters that were either pressed into the person's palm or were freely explored. He found that active touch achieved a 95% success rate at identifying cutter shape whilst a passive sensation upon the person's skin only achieved 29% correct identification (chance being 16.7% correct). In a similar experiment, Heller [25] limited the time participants had to explore objects through touch and found that to match the performance of 5 s of active touch, participants needed over 30 s of passive stimulation. Along similar lines, Klatzky et al. [33] asked people to identify one hundred common objects through active touch and found that almost all of them could be identified within a few seconds.

Haptics comprises all the elements of passive touch (mechanoreceptor sensation) plus information from the kinaesthetic system and other higher order cognitive processes such as memory and attentional processes. When exploring an object Lederman and Klatzky [36] observed that participants were using highly stereotypical hand movements and over a series of studies, they categorised and explored each movement. These stereotyped movements were termed 'Exploratory Procedures' (EPs), and each EP is specific to the task for the observer. Examples of such EPs include 'lateral motion', which is used to explore an object's surface texture; 'contour following', which is used for shape perception and 'part motion test' where the range and resistance of moving parts are determined.

Lederman and Klatzky argued that when confronted with an object to be recognised, a person will first employ a general EP that is fast and provides adequate information across multiple dimensions of the object. This information is then used to decide which of the more task specific set of EPs to use and to test hypotheses about the identity of the object. Thus, a cycle of EP selection and re-sampling repeats until the object is identified or sufficient knowledge has been extracted in order to recognise the object. Lederman and Klatzky [37] termed this a two-process sequence of haptic exploration.

2.3. Similarities and differences between vision and haptics

Comparing vision with touch, a distinction can be made between the nature of encoding information about an object and the type of information that can be encoded. In terms of the nature of encoding information, vision and touch differ in many ways. Two of the more of the obvious differences include the rate and the range of sampling. Vision can readily access information about an object's properties, whereas for touch, an EP must be initiated to gain such information. Even though multiple features of an object can be sampled simultaneously with a broad enough EP, the haptic system can only explore a set of objects one at a time through a series of 'successive impressions' [20]. Furthermore, the haptic system typically works within a limited range around the body, usually within peripersonal space, and can only interact with objects that are within reachable distance. For vision, an entire scene can be viewed at the same time, across a much larger scale. Finally, some features of an object can only be perceived by one modality, usually referred to as a modality encoding bias [38]. For example colour can only be directly perceived visually whilst object mass, surface friction, hardness and temperature can only be perceived directly through touch (although it may be possible to infer these properties through another modality; for example, steam rising from the glass would suggest the contents are hot). These cross-modal differences in the timing of information pick-up and the modality encoding range may affect cross-modal performance.

Vision and haptics are similar however in terms of being able to only sample in detail one object at a time, although vision can often achieve this in a fraction of the time it would take touch [36]. Both modalities are capable of measuring object 'macrogeometric features' (a term coined by [61]) referring to, for example, object orientation, size, and gross shape, although the visual system seems to be more involved in the perception of macrogeometric features [38] even in the haptic perception of these features [61]. Zangaladze et al. found, for example, that tactile orientation discrimination in normally sighted individuals was impaired when focal transcranial magnetic stimulation (TMS) was applied to the visual cortex effectively disrupting visual processing. They argue that it is because we generally rely on vision to perceive macrogeometric features such as orientation, that it has an influence on the haptic perception of that feature. To a certain extent, both modalities are able to measure 'microgeometric' features such as material differences including surface texture [38], although texture

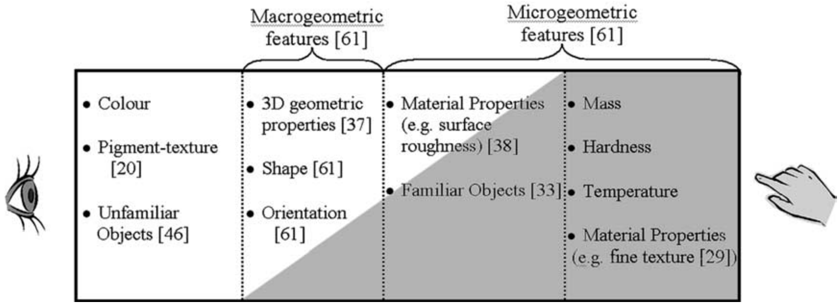
Macrogeometric features [61]
Microgeometric features [61]
• Colour
• Pigment-texture [20]
• Unfamiliar Objects [46]
• 3D geometric properties [37]
• Shape [61]
• Orientation [61]
• Material Properties (e.g. surface roughness) [38]
• Familiar Objects [33]
• Mass
• Hardness
• Temperature
• Material Properties (e.g. fine texture [29])

Fig. 2. Information about an object can either be modality specific or can be shared across modalities. In our illustration here, we have mapped task dependent modality biases. For example, only vision can perceive colour information (left of figure) whereas only touch can perceive mass. Other features, such as orientation, can be perceived by both modalities. However, each modality can differentially influence the perception of these features. For example, Zangaladze et al. [61] suggest that orientation perception is dominated by vision. On the other hand Lederman et al. [38] argue that the perception of some material properties is dominated by touch. Here we characterise the relative influence of each modality on the perception of different types of object features based on some studies in the literature. Note however that our figure here is by no means comprehensive and acts mainly as a guide for modality biases in perception.

perception is more precise in touch than in vision [29], suggesting that tactile perception may influence the visual perception of surface texture. In Fig. 2 we illustrate a number of encoding biases and similarities between vision and haptics for the purposes of recognising an object.

The differences or modality encoding biases between haptics and vision need to be taken into account in any investigation of cross-modal integration. Several studies have controlled for such differences. For example, Loomis et al. [40] attempted to equate haptic and visual learning of stimuli by allowing participants to look at the stimuli successively through a small aperture during visual learning. In other studies, visual and haptic performance was rendered equivalent by allowing participants more time during haptic learning of objects (e.g. [47]).

Despite these differences, the presentation of a single object to vision and touch simultaneously will yield a common subset of information that can be easily shared across these modalities. When cross-modal encoded information about an object is both spatially and temporally congruent it is suggested that this information is automatically integrated (e.g. see [4] for integration between audition and vision). Such integration allows for the creation of a rich cross-modal representation of the object in memory. It is unclear, however, how information is integrated across modalities and when integration is achieved. Some recent studies on cortical areas involved in cross-modal perception suggest some possible solutions as to how integration occurs.

3. Cross-modal integration in the cortex

Broadly speaking, two accounts of multisensory integration are discussed within the neurological literature. The first view was evident in the 1960s to 1980s where vision and touch were seen as structurally and functionally independent systems (e.g. [15]) that broadly adhered to Fodor's rules of modularity [16]. Accordingly, a large volume of literature is dedicated to processing in each of these modalities separately. Cross-modal integration was not considered to occur directly, but through an intermediatory or translation process such as through imagery. Recent findings on neuronal or cortical plasticity, however, have challenged this notion of modularity: it is now accepted that cortical areas traditionally considered as sensory specific are not necessarily so and can, for example, be recruited by other areas in the case of sensory deprivation [50].

The second, more recent, account suggests that dedicated multisensory areas exist within the brain to process and perhaps store information from two or more modalities. Researchers have used a variety of techniques and paradigms to identify candidate areas within the brain involved in shape recognition. For example, Banati et al. [2] conducted a PET study during a shape matching task and found that anterior cingulate, inferior parietal lobules and claustrum areas were selectively activated during cross-modal matching. Using fMRI, Amedi et al. [1], investigated cortical areas involved in cross-modal, familiar object recognition. They found that the lateral occipital complex (LOC), a cortical area well known to be involved in visual object recognition, was also active during haptic recognition of familiar objects. This activation was reduced when visual imagery was controlled for, suggesting that the LOC is directly implicated in recognition. They suggested that the LOC is a candidate cortical area for the convergence of multisensory information about objects. Another recent fMRI study involving haptic to visual priming of novel objects also provided evidence for the role of area LOC in cross-modal object recognition [28].

The question still remains however as to the architecture of these multisensory areas at the neuronal level.

For example, area LOC may be populated by neurons that are each selective to sensory information from one modality only. Therefore, areas such as the LOC are multisensory only because they are populated by multiple, sensory specific neurons. Others suggest the claustrum as the candidate structural area for multisensory integration [14,23]. Hadjikhani and Roland [23] propose that multisensory cortical areas perhaps act as translator mechanisms between the modalities rather than within-modal representation storage areas. On the other hand, individual neurons in LOC may be multisensory and selective to particular features irrespective of modality. Such multisensory neurons have been found in sub-cortical structures such as the superior colliculus [49,56] and in cortical areas such as parietal areas involved in representing extra-personal space [22]. It is feasible, therefore, that there exist multisensory neurons selectively tuned to features of an object encoded through any modality. To our knowledge, however, there are no studies reporting the existence of such neurons.

More recent studies have reported that primary sensory areas can also be involved in cross-modal tasks. For example, Zangaladze et al. [61] found that applying transcranial magnetic stimulation (TMS) to the occipital cortex, an area previously thought to be dedicated for visual processing, hindered performance on a haptic grating orientation task. Conversely, visual tasks involving stimuli with tactile qualities have been shown to activate somatosensory cortex. Zhou and Fuster [62] recorded activity in individual neurons in monkey somatosensory cortex and found that visual stimuli that are behaviourally associated with tactile information activated a number of these neurons (in this case, bars with horizontal or vertical gratings, perceivable by the haptic system as texture). The task required the monkey to remember the grating orientation for a short period and then make a decision about which of two haptically presented gratings they had just seen. Zhou and Fuster suggest that neurons in somatosensory cortex are involved in the cross-modal short-term memory of object features.

Instead of proposing that cross-modal information is either coded by multisensory neurons or cortical areas populated with different sensory specific neurons, Meredith [44] proposes that sensory convergence might be considered as a continuum (see Fig. 3 for an illustration). He argues that at one extreme of the continuum convergence is areal where sensory specific neurons from several modalities converge in the same cortical area. At the other end of the continuum convergence is neuronal where a single neuron is able to respond to inputs from one or both sensory streams. He argues that many cortical systems may lie somewhere along this continuum coding both sensory specific and multisensory inputs.

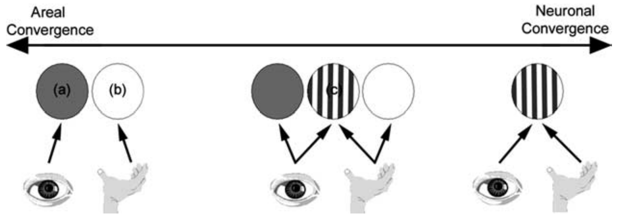

Fig. 3. An illustration of areal to neuronal convergence continuum based on Meredith [44]. Three sets of neurons (○) are evident in the above figure and each receives input from touch and/or vision. In our illustration, those neurons receiving input from vision are shaded grey (a) whilst the neurons receiving input from touch are white (b). Hashed neurons receive input from both modalities (c). Areal convergence is where different modality specific neurons populate the same cortical or subcortical area (the left of the figure) and neuronal convergence is where a single neuron receives input from two or more modalities (the right of the figure). Meredith argues that most multisensory areas are characterised by both areal and neuronal convergence as illustrated in the centre of this continuum.

Interestingly, a number of behavioural effects can be predicted for each of these extreme accounts. One prediction is that the efficiency in accessing information stored in multisensory areas should be independent of the sampling modality. On the other hand, cross-modal performance should be reduced if uni-modal areas need to communicate information between each other because accessing of information across modalities may require a translation process. Such a process may result in a cost in performance. We discuss studies investigating these issues in the following sections.

Observations of neural activation alone might not reveal phenomenological experiences or the efficiency of the mechanism involved in performing a cross-modal recognition task. Behavioural studies can better reveal how information is shared across modalities. For example, neuronal activation would not reveal the nature of the subjective experience in the individual nor whether the person is experiencing an illusory percept or a veridical percept. The next section includes a discussion of studies conducted in our laboratory on the behavioural correlates of within- and cross-modal recognition.

4. Experimental studies on cross-modal recognition

It is clear that every-day recognition tasks (i.e. those outside the laboratory) involve the pick-up of information from several senses, giving rise to the question of whether object constancy is solved through cross-modal processing. In the following sections, we review recent experiments investigating the role of cross-modal processing on object recognition. Since recognition in the real world is not strictly confined to single objects, we also report on the role of cross-modal processing in scene perception.

4.1. Cross-modal recognition of single objects

In a recent study, Newell et al. [47] reported that vision and haptics can complement each other in forming a representation of a three-dimensional object. The authors were particularly interested in investigating whether combining information across the modalities would help solve some of the problems affecting object constancy, particularly the problem for recognition of changing viewpoints (e.g. [10,46]). In our experiments we measured the effects of viewpoint on visual, haptic and cross-modal recognition. We used a set of unfamiliar objects made from six identical red Lego™ bricks. All objects were presented in a fixed position in front of the participant. Each object had a different configuration of these bricks, and hence we controlled for differences in size, texture and colour, which may aid recognition performance in one modality only (see Fig. 4 for an illustration). Participants performed a recognition memory task in all the experiments. First, they were required to learn four target objects in a sequential order either visually (for 30 s each) or haptically (for 1 min each). A pilot study revealed that haptic exploration time needed to be longer than visual exploration time in order to achieve equivalent performance across these modalities. The objects were placed behind a curtain during the haptic condition and participants placed their hands underneath the curtain to touch the objects. We gave no explicit instructions on how to learn the objects except that the object should not be moved. A test session followed each learning session. During the test, objects were presented sequentially and the four target objects had to be recognised from four similar, non-targets. Each target object was presented in either the same orientation as learning (0°), or rotated by 180°, and in the same modality as learning or in the other modality.

First, we found that within-modality recognition was view-dependent. Therefore, the literature on view-dependent effects in visual object recognition was replicated whilst a new finding of view-dependency in haptic object recognition was also presented. Although effects of orientation in tactile perception, such as the mental rotation of abstract forms [6], have previously been reported in the literature, the effect of orientation in haptic object recognition was not. Interestingly, in the cross-modal condition the effects of view-dependency were reversed (see Fig. 5). In other words, cross-modal recognition was better when there was a change in orientation between learning and test, but only when there was an interchange between the front and the back of the objects and not with left/right changes. These findings suggest that both the visual and haptic systems code view-specific representations of objects, and that each system has its own preferential 'view' of an object: for the visual system, it was the surface of the object facing the observer whereas for the haptic system, it was the back surface of the object (with respect to the observer). The finding that the haptic system preferred the back surface of objects was determined in a separate experiment in the Newell et al. study. Here they measured haptic recognition to either the front or back surface by occluding information from the remaining surfaces in each condition. Recognition performance was better to the 'back' surface. We surmise that this 'back' surface is preferential for the haptic system mainly because of the position of the hands in relation to the body which in turn allows for better shape exploration by the fingers. Representations of these 'best' views from vision and haptics can then be easily shared across modalities when they are in correspondence (i.e. when the hands explore a previously viewed surface and vice versa).

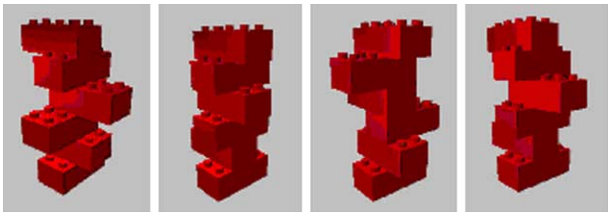

Fig. 4. An illustration of the type of objects used in our haptic and visual object recognition experiments.

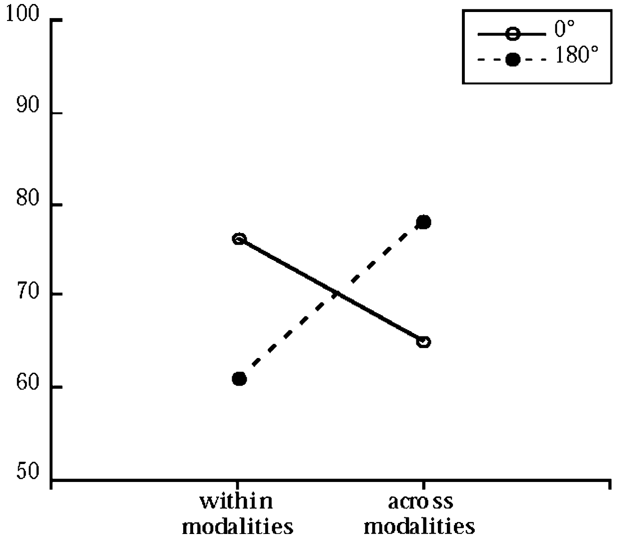

Fig. 5. Plot showing the results of our single object recognition experiments across changes in viewpoint between learning and test involving an interchange of the front and back surfaces (180°) or no viewpoint change (0°). A change in viewpoint caused a decrease in percent correct responses within modalities but an increase in percent correct responses across modalities. See text for an explanation of our findings.

On a similar note, an interesting study was recently reported on haptic face recognition. Kilgour and Lederman [32] tested haptic and cross-modal matching performance of real faces and face masks. They found that for some face-stimuli groups visual to haptic face matching was more accurate than haptic to visual face matching. Haptic matching performance did not improve by the presence of vision during learning, leading to the suggestion that a visual code did not easily transfer to haptics for matching purposes. However, based on the findings in the Newell et al. [47] study, there may be another possible explanation for their cross-modal results: during haptic learning the face stimuli may not have been presented in an optimal orientation with respect to the observer, to allow for efficient cross-modal matching. If, for example, the face stimuli were positioned such that they are facing away from the participant, and the participant's hands explored the face from that position, i.e. a more natural position for the hands as argued earlier, then cross-modal performance might improve. Clearly, more research is required to test the generalisability of the 'haptic view' claim.

Although other studies have also emphasised the role of finger exploration for efficient haptic object recognition [34], the Newell et al. [47] study further proposes that the haptic representation of an object can be shared with the visual system by matching across surfaces. In order for cross-modal access to an object's representation to occur, however, the same surface should be perceived by both modalities. Under conditions where cross-modal surface correspondence is not optimal, for example if the object can be freely moved, then we would predict poor cross-modal performance in this case because the 'best' surfaces are unlikely to be properly matched and integrated. This prediction was recently tested in a cross-modal recognition study under conditions of unconstrained viewing [13]. The same stimuli and design as in the Newell et al. [47] study were used, except that in this study, participants were allowed to palpate freely the objects in the haptic modality. For the visual condition, each object was placed into a transparent Perspex ball. In this way, the participants could freely view each object without actually touching the object. Both within-modal and cross-modal recognition was measured.

When viewing or palpation is unconstrained, cross-modal recognition was worse than within-modal recognition (see Fig. 6). This finding fits with the prediction that efficient cross-modal recognition depends upon spatially constraining the surfaces of objects in order to match the best 'views' across modalities. In other words, spatial integration of the surfaces of the objects is the key to efficient cross-modal recognition. We might ask, however, how cross-modal recognition is affected when both modalities are used simultaneously during learning, i.e. when the object is actively touched in the presence of vision, relative to when the object is either visually or haptically explored during learning. Here we would predict that cross-modal learning would enhance recognition performance because all surfaces are necessarily in correspondence during learning. An interesting study reported recently by Harman et al. [24] also suggests a role for surface correspondence across modalities in object recognition. They found that when an observer could actively move a virtual image of an object on a screen recognition performance was better than when active movement was not conducted and the observer passively viewed the same sequence of object views. Therefore, manipulation enhanced visual recognition performance. Haptic perception per se was not tested in the Harman et al. study, therefore, the question still remains as to whether cross-modal active learning (i.e. seeing and feeling the object at the same time) can enhance recognition relative to active visual or haptic learning alone. We recently tested this notion and found that indeed cross-modal learning of novel objects promoted better recognition performance than either visual or haptic learning alone [13].

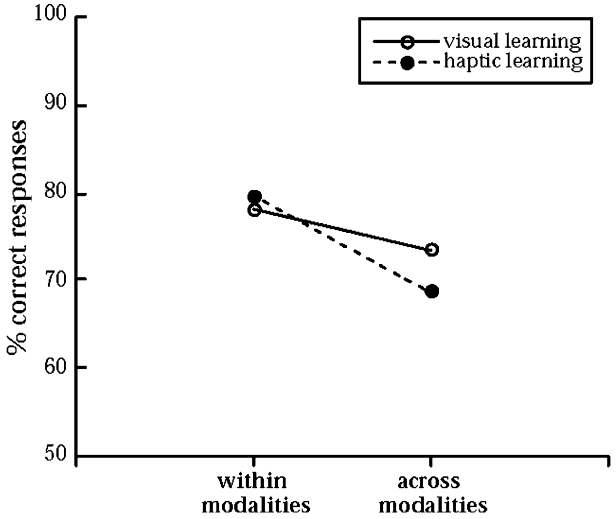

Fig. 6. Plot showing the effect of within-modal and cross-modal recognition of objects under unconstrained 'viewing' conditions, i.e. the objects could be freely moved for visual or haptic sampling. We found that cross-modal recognition performance was worse than within-modal performance indicating that spatial correspondence is necessary for cross-modal object recognition.

4.2. Cross-modal recognition of multiple objects

The simultaneous recognition of more than one object, for example a scene of objects, clearly presents a different problem for cross-modal recognition because of the larger spatial scales involved. It is unlikely, for example, that fast scene recognition would involve the type of exploratory hand movements typical for single object recognition [36]. Spatial integration for scene recognition involves integrating across objects over a possibly large scale thus affecting larger differences between visual and haptic encoding. There may be a sequential nature to global scene exploration within the haptic domain [26] (i.e. the hands move to one object at a time), which would mean that the spatial structure of the scene is only available when information about the entire content of the scene is integrated. Visual scene recognition, on the other hand, does not necessarily involve the same serial processes. In large-scale scenes, however, spatial integration would be required across eye fixations [27,51]. Generally though, the contents of a scene can be rapidly perceived by the eye [3]. Haptic space, on the other hand, is limited to peripersonal space, although it does have some flexibility such as when tools are wielded [35]. Furthermore, the representation of haptic space has been found to be distorted, i.e. non-Euclidean compared to the representation of visual space which is considered to be Euclidean [30].

Recent research has shown that our visual memory for scenes is sensitive to changes in orientation with respect to the observer [7,55]. In the Diwadkar and McNamara study, participants were first required to study a real scene of 6 objects for 30 s. Their memory for this scene was then tested using images of the target scene from different angles against similar distractor scenes (i.e. novel configurations of the same objects). They found that scene recognition performance was a function of angular distance to the original view (Experiment 1) or different training views (Experiment 2). Simons and Wang [55] also tested the effects of orientation on the recognition of real scenes of objects placed on a table. They found that 47° changes in the orientation of a scene of objects between learning and test significantly reduced recognition performance. Interestingly, scene recognition performance was not just dependent on the retinal projection of the scene: no effects of orientation were found when the observer could move themselves to a new viewing position 47° away from the learning position. This again indicates the importance of active perception. These studies, therefore, suggest that visual spatial memory performance is similar to single object recognition, at least in terms of orientation-dependency. But what was not known is how similar haptic spatial memory is to visuo-spatial memory.

By investigating cross-modal recognition performance we may be able to provide a clearer understanding of the nature of the encoded information within each modality and how this information is shared in order to recognise scenes of objects. In our study, we tested observer's ability to recognise the spatial layout of objects in a scene both uni-modally (either haptically or visually) or across modalities [48]. Specifically, our two experimental questions were (a) is cross-modal scene recognition efficient relative to within-modal recognition and (b) is cross-modal recognition affected by changes in the orientation of the scene?

The stimulus set of objects used included 15 wooden shapes of familiar objects, seven of which were randomly placed on a rotatable platform in any one trial (see Fig. 7 for an illustration of a typical scene shown in both 0° and 60° rotations). We attempted to provide equal accessibility to object properties across both modalities by allowing participants 10 s to view the scene and 1 min for haptic exploration during learning. These presentation times were determined in a pilot study and give equivalent within modality recognition performance across vision and haptics. After learning, the position of two objects was exchanged whilst keeping the orientation of the individual object constant and participants had to identify the swapped objects at test. Thus within a trial the same object positions were occupied between learning and test although two objects were swapped at test. We maintained configural information of the object scenes for two reasons: first, at least for visual processing, we did not want participants extracting a global shape of the scene in order to do the task. Second, we wanted to ensure that haptic memory was based on touch alone and not on proprioceptive cues such as body movement or body-centred posture due to changes in the global configuration [45]. Testing occurred either within or across modalities and under 0° or 60° rotation of the scene.

We found that recognition performance was impaired by a change in modalities at test. We also found that both within- and cross-modal scene recognition was sensitive to orientation changes with respect to the ob-

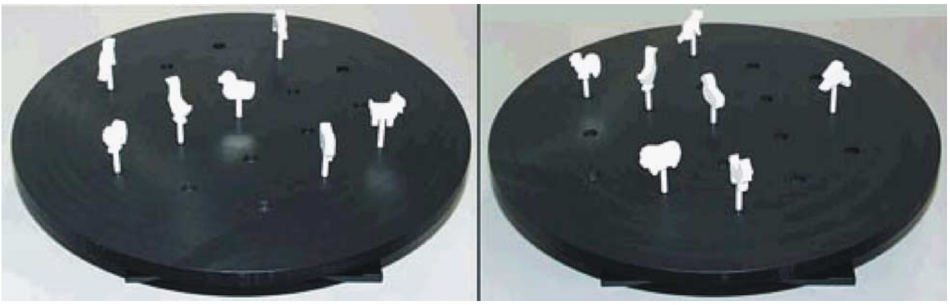

Fig. 7. An illustration of a typical target scene used in our haptic and visual multiple object recognition experiments. The same scene is depicted from a 0° angle on the left and from 60° rotation on the right.

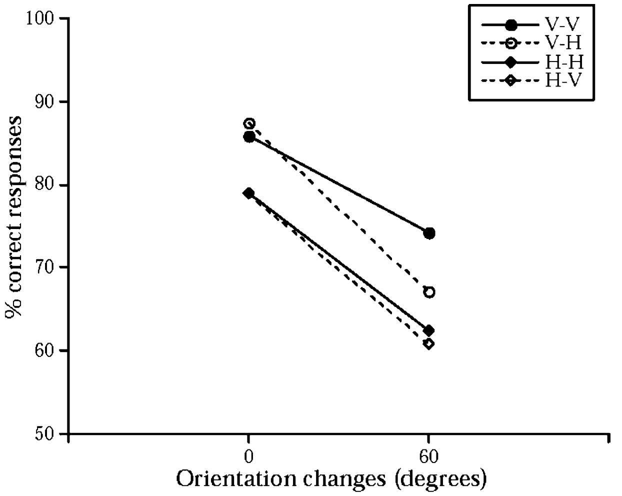

Fig. 8. Plot showing the effects of orientation across within-modal (V–V and H–H) and cross-modal (V–H and H–V) scene recognition conditions. Changes of 0° and 60° refer to changes in the orientation of the scene between learning and test. A cost in recognition performance was found with changes in orientation of 60° and with changes in modality.

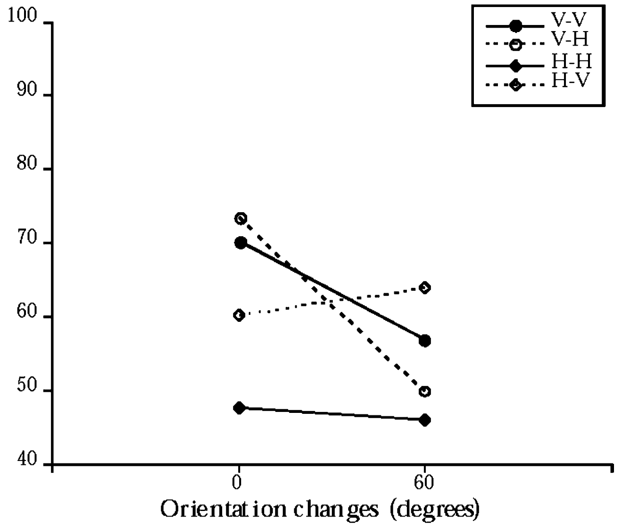

Fig. 9. Plot showing the effects of a verbal interpolation task on within-modal (V–V and H–H) and cross-modal (V–H and H–V) scene recognition conditions with changes in orientation. Changes of 0° and 60° refer to changes in the orientation of the scene between learning and test. The within-modal haptic condition (H–H) was affected by the interpolation task relative to conditions involving visual learning (V–V and V–H) or visual testing (H–V).

server (see Fig. 8). This result indicates that vision and haptics can share a common, orientation sensitive code for scene perception.

The relative cost in cross-modal recognition suggests a difference in encoding of information between vision and haptics. One possible difference is that visual information can be encoded holistically over large scales whereas haptic encoding is generally serial. However we also noticed that some participants recalled the objects' names. We therefore decided to test the role of verbal mediation in cross-modal recognition performance in a separate experiment. We repeated our experiment described above but now required participants to perform a verbal interpolation task (i.e. generate lists of words aloud beginning with a random letter) between learning and test. Again we found that cross-modal recognition was less efficient than within-modal performance. More pertinently, verbal interference produced more errors in all conditions involving haptic perception, relative to the vision–vision condition (see Fig. 9). Furthermore, the effect of orientation remained for within-modal but not for cross-modal recognition, suggesting a demand on central processes which made the task more difficult. The relatively poor transfer from haptics with verbal interference suggested a modality encoding bias, with evidence for a role of verbal recoding in haptic perception.

Other studies have also shown an effect of verbal interference in haptic memory tasks. When two sequences of tactile stimuli were presented to the fingers, Mahrer and Miles [41] reported that matching performance was affected by the presence of a verbal interpolation task during the delay between presentations but not by the presence of a tactile interference task. This result suggests a role for verbal rehearsal in the memory of tactile sequences. However, their results also indicate the role of visuo-spatial processing: when the task required matching the tactile sequence to a visual image of a hand then performance improved. Mahrer and Miles therefore argue that tactile memory can exploit both visuo-spatial and verbal processes for recall purposes. Our data also indicate a role for both a verbal recoding in haptic learning, together with a more visuo-spatial encoding for the purposes of cross-modal recognition. Furthermore, these data suggest that both visual and haptic encoding involves a mapping of spatial relations between objects into, perhaps, a body-centred reference frame [9] thus accounting for the effect of orientation within and across modalities.

Taken together, our findings from single object and multiple object recognition tasks indicate that orientation dependency is a common characteristic of different sensory systems. For cross-modal recognition the findings from our studies suggest that efficient performance is dependent on the nature of encoding shapes across modalities. For example, visual and haptic recognition of single objects was view-dependent and efficient cross-modal performance was based on matching across corresponding 'views' or surfaces of these objects. This finding suggests that it is a common code (i.e. surface related), coupled with common processing (i.e. view dependence) which allows for cross-modal recognition. Thus, when there was a possible discrepancy of encoding across modalities as in the scene perception studies (e.g. spatial encoding for vision versus serial encoding, or verbal recoding, for haptics), cross-modal recognition was less efficient. This reduction in efficiency may be due to a translation process of information from one modality into a format congruent with the other modality. We are currently investigating these possible mechanisms underlying the cross-modal recognition of objects and scenes.

5. Factors affecting cross-modal integration

A number of factors can affect the nature and extent of cross-modal integration. Chiefly amongst these factors are (a) the demands of the task, (b) encoding conditions and (c) spatial and temporal congruency of object information. These factors will be discussed in the following sections.

5.1. Task effects

As already discussed, modality specific attributes lead to an automatic capture of certain object properties. If the task itself specifies the object properties that are predominantly sampled by one modality over the other, then a degree of sensory capture will be evident. Interestingly, Lederman and Klatzky [36] observed that sometimes when vision and touch were used together, hand exploration could be dramatically reduced. The predominance in a situation of one modality over the other not only may prevent integration, but also perhaps prevent encoding in the 'weaker' modality.

It was recently reported that the extent to which information from one modality is used over another depends on the reliability of that information within each modality [12]. Previously, Rock and Victor [53] argued that for shape perception vision can dominate touch, thus coining the term 'visual capture'. However, more recent studies have suggested that 'capture' is an extreme form of cross-modal processing and that decisions are generally based on the reliability of information in each modality. For example, Ernst and Banks [12] systematically adjusted the reliability of visual information in a visual-haptic height judgement task. Using a cross-modal cue conflict paradigm, Ernst and Banks decreased the reliability of the visual information and found that haptic information increasingly influenced the participant's percept. Thus, the integration of visual and haptic information is achieved in a statistically optimum way with the use of sensory information from either vision or haptics being dependent on the estimated reliability of that information.

The familiarity of the task and stimuli may also affect integration. When learning to drive, for example, initially each cross-modal action is performed explicitly until automatic co-ordination between vision and touch is achieved. At first, changing into third gear requires a glance at the gear stick to find out exactly where it is, a press on the clutch with your foot and then a poorly timed, clumsy movement of the gear stick. As much information as possible is gathered about the task of changing gear through touch and vision (and through audition when the car screams its protests) but through practice, such overt cross-modal integration is not needed and the gear can be changed through feel alone. Familiarisation can lead to the optimisation of information encoding such that other sensory information is no longer explicitly encoded but can be retrieved from memory for task completion.

5.2. Attentional effects and feature integration

The role of attention on feature integration in vision has been extensively studied over the past few decades. In 1986, Treisman [59] proposed that attention is necessary in order to bind, or integrate across features of an object. She used a visual search paradigm to illustrate her argument. If a target object is defined by a unique feature relative to non-target objects (e.g. a green X amongst red Xs), attention is not required in order to identify or locate the object. This is the so-called pop-out effect, and is characterised by a flat slope of response times to locating the target plotted against number of non-targets in a display. If the target is specified by a conjunction of features that can also belong to non-targets (e.g. searching for a green X among green Ys and red Xs), then serial attention is required on each object in a display in order to locate the target. Thus response times would increase with an increase in the number of non-targets in a display. Although, as far as we are aware, this idea has not yet been tested in a cross-modal context, we might assume that the same principle would extend to a cross-modal search task. If a target object is defined by a conjunction of cross-modal features then we would expect that locating or identifying the target would require serial attentional processing in order to integrate features of the object across modalities. Driver [8] noted that some cross-modal binding between vision and audition for the purpose of speech localisation can occur without the need for attention. However, we do not yet know how attention affects our perception of cross-modal events that are less familiar than speech recognition or localisation.

5.3. Encoding effects

For integration to occur, the same object must be experienced by both vision and touch at the same time. Optimum integration depends on the type of object being sampled and the amount of information encoded. For example, Calvert et al. [4] argue that for stimuli with low information content (e.g. a light flash and a sound burst) then temporal congruency would be sufficient for optimal binding to occur. On the other hand, stimuli with high-level information content (i.e. more complex objects) may have to rely on a combination of temporal and spatial congruency for optimum integration.

Two spatially independent events can be integrated if a high degree of temporal congruency is evident between them. The effect of such integration can often be illusory. Perhaps the best known examples of cross-modal illusions concern the interaction between audition and vision, for example the McGurk or ventriloquism effects [8,42]. However, the introduction of a delay between the presentation of a stimulus from one modality to the other would likely affect integration. A delay would require the first presentation to be held in memory and this would be subject to decay effects and possibly interference effects [21,39,57]. What is not yet known, however, is the extent of the time window within which integration occurs, although this is likely to be dependent on the decay rate of information across modalities.

Spatial congruency across modalities can either enhance known information about an object or degrade it. When information across modalities is spatially congruent then cross-modal recognition can be enhanced [13]. Reales and Ballestersos [52] found that when an object was presented in one modality (vision or haptics) the recognition of that same shape in the other modality was facilitated. Therefore, spatial congruency allowed for cross-modal priming. Spatial congruency can also result in cross-modal illusions. For example, touch (particularly in judging the dimensions of an object for grasping) can be affected by visual illusions such as the Ebbinghaus illusion [17]. These findings suggest that spatially congruent information is processed in a similar way across modalities. On the other hand, low spatial congruency can lead to a degradation of an object's representation in memory [31,39,41]. For example, Kerzel [31] found that participant's memory for the velocity of a visual stimulus was affected by different movement of the participant's own arm. A simultaneous but relatively slower arm movement resulted in a memory of a slightly slower visual stimulus than the veridical speed of the visual stimulus, and the opposite effect was found with a faster arm movement.

6. Conclusions

In summary, the aim of this paper was to explore cross-modal object and scene recognition and to determine the factors which influence integration across these modalities. We first reviewed the literature on object recognition within vision and touch separately. A number of distinctions were made between the way in which vision and haptics encode and represent object information based on the object features each modality can encode. We then proposed two different accounts of how cross-modal information might be integrated in the cortex. In the first account, integration occurs when unimodal areas communicate indirectly with each other through an intermediatory or translation process. In the second it is proposed that multisensory neurons are tuned to information from all modalities. Although, through recent neuroimaging studies a number of cortical structures have been implicated in cross-modal object recognition it is not yet clear whether these areas are involved in translating information from one modality to the next or whether they are truly multisensory areas at the neuronal level.

Our own behavioural investigations suggest that cross-modal recognition is dependent on a common code and a common form of processing between the two modalities: for single object recognition we found that information about an object's surfaces is encoded by both the visual and haptic system. Cross-modal recognition in this case is dependent on the direct correspondence of surface information across the modalities. For scene recognition we also found that each modality was sensitive to the general orientation of the scene with respect to the observer. However, cross-modal recognition was less efficient than within-modal recognition. We suggest that poor cross-modal transfer in scene recognition may be due to the differences in the encoding of large-scale items across the modalities; visual encoding is largely holistic over large scales whereas haptics is based more on the serial encoding of objects. Finally, factors necessary for the optimal integration of cross-modal integration were briefly discussed, including task demands, attentional and encoding effects.

Vision and touch are very different information capture systems that measure a range of object attributes. Through integration of information across the senses we are able to create a much richer representation of our world. We have discussed behavioural and neurological evidence that suggest mechanisms for how cross-modal integration might occur across vision and haptics. Clearly, however, many questions remain outstanding before we can provide a better understanding of how, where and when such rich representations are developed that allow for optimal recognition performance.

Acknowledgements

The research reported in this article was funded by, Trinity College Dublin Research Fund, HEA Fund for collaborative research between Irish Third-Level Institutions and Media Lab Europe, and the Max Planck Society, Germany. We thank Síle O'Modhráin and Charles Spence for comments on an earlier draft.

References

[1] A. Amedi, R. Malach, T. Hendler, S. Peled, E. Zohary, Visuo-haptic object-related activation in the vental visual pathway, Nat. Neurosci. 4 (2001) 324–330.

[2] R.B. Banati, G.W. Goerres, C. Tjoa, J.P. Aggleton, P. Grasby, The functional anatomy of visual-tactile integration in man: a study using positron emission tomography, Neuropsychologia 38 (2000) 115–124.

[3] I. Biederman, J.C. Rabinowitz, A.L. Glass, E.W. Stacy, On the information extracted from a glance at a scene, J. Exp. Psych. 103 (1974) 597–600.

[4] G.A. Calvert, M.J. Brammer, S.D. Iversen, Cross-modal identification, Trends Cognit. Sci. 2 (1998) 247–253.

[5] J.C. Craig, G.B. Rollman, Somesthesis, Ann. Rev. Psych. 50 (1999) 305–331.

[6] A. Dellantonio, F. Spagnolo, Mental rotation of tactual stimuli, Acta Psych. 73 (1990) 245–257.

[7] V.A. Diwadkar, T.P. McNamara, Viewpoint dependence in scene recognition, Psych. Sci. 8 (1997) 302–307.

[8] J. Driver, Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading, Nature 381 (1996) 66–68.

[9] R.D. Easton, M.J. Sholl, Object-array structure, frames of reference, and retrieval of spatial knowledge, J. Exp. Psych. Learn. Mem. Cog. 21 (1995) 483–500.

[10] S. Edelman, H.H. Bülthoff, Orientation dependence in the recognition of familiar and novel views of 3-dimensional objects, Vision Res. 32 (1992) 2385–2400.

[11] S. Edelman, Representation and Recognition in Vision, Bradford Books, MIT Press, Cambridge, MA, USA, 1999.

[12] M.O. Ernst, M.S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion, Nature 415 (2002) 429–433.

[13] M.O. Ernst, H.H. Bülthoff, F.N. Newell, Visual and haptic recognition of actively explored objects, in preparation.

[14] G. Ettlinger, W.A. Wilson, Cross-modal performance: behavioural processes, phylogenetic considerations and neural mechanisms, Behav. Brain. Res. 40 (1990) 169–192.

[15] D.J. Felleman, D.C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex, Cereb. Cortex 1 (1991) 1–47.

[16] J.A. Fodor, The Modularity of Mind, Bradford Books, MIT Press, Cambridge, MA, USA, 1983.

[17] V.H. Franz, K.R. Gegenfurtner, H.H. Bülthoff, M. Fahle, Grasping visual illusions: no evidence for a dissociation between perception and action, Psych. Sci. 11 (2000) 20–25.

[18] W.W. Gaver, What in the world do we hear? An ecological approach to auditory event perception, Ecol. Psych. 5 (1993) 1–29.

[19] W.W. Gaver, How do we hear in the world: explorations in ecological acoustics, Ecol. Psych. 5 (1993) 285–313.

[20] J.J. Gibson, Observations on active touch, Psych. Rev. 69 (1962) 477–491.

[21] E.Q. Gilson, A.D. Baddeley, Tactile short-term memory, Quart. J. Exp. Psych. 21 (1969) 180–184.

[22] M.S.A. Graziano, C.G. Gross, The representation of extrapersonal space: a possible role for bimodal, visual-tactile neurons, in: M.S. Gazzaniga (Ed.), The Cognitive Neurosciences, vol. XIV, The MIT Press, Cambridge, MA, USA, 1995, 1447 pp.

[23] N. Hadjikhani, P.E. Roland, Cross-modal transfer of information between the tactile and the visual representations in the human brain: a positron emission tomographic study, J. Neurosci. 18 (1998) 1072–1084.

[24] K.L. Harman, G.K. Humphrey, M.A. Goodale, Active manual control of object views facilitates visual recognition, Curr. Biol. 9 (1999) 1315–1318.

[25] M.A. Heller, Active and passive touch: the influence of exploration time on form recognition, J. Gen. Psych. 110 (1984) 243–249.

[26] M.A. Heller, S. Clyburn, Global and local processing in haptic perception of form, Bull. Psycohonomic Soc. 31 (6) (1993) 574–576.

[27] J.M. Henderson, A. Hollingworth, High level scene perception, Ann. Rev. Psych. 50 (1999) 243–271.

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

[29] R.S. Johansson, R.H. LaMotte, Tactile detection thresholds for a single asperity on an otherwise smooth surface, Somatosens. Res. 1 (1983) 21–31.

[30] A.M.L. Kappers, J.J. Koenderink, Haptic perception of spatial relations, Perception 28 (1999) 781–795.

[31] D. Kerzel, Visual short-term memory is influenced by haptic perception, J. Exp. Psych. Learn. Mem. Cog. 27 (2001) 1101–1109.

[32] A.R. Kilgour, S.J. Lederman, Face recognition by hand, Percept. Psychophys. 64 (2002) 339–352.

[33] R.L. Klatzky, S.J. Lederman, V. Metzger, Identifying objects by touch: an 'expert system', Percept. Psychophys. 37 (1985) 299–302.

[34] R.L. Klatzky, J.M. Loomis, S.J. Lederman, H. Wake, N. Fujita, Haptic identification of objects and their depictions, Percept. Psychophys. 52 (1993) 170–178.

[35] E. Làdavas, Functional and dynamic properties of visual peripersonal space, Trends Cognit. Sci. 6 (2002) 17–22.

[36] S.J. Lederman, R.L. Klatzky, Hand movements: a window into haptic object recognition, Cog. Psych. 19 (1987) 342–368.

[37] S.J. Lederman, R.L. Klatzky, Haptic classification of common objects: knowledge driven exploration, Cog. Psych. 22 (1990) 421–459.

[38] S.J. Lederman, C. Summers, R.L. Klatzky, Cognitive salience of haptic object properties: role of modality-encoding bias, Perception 25 (1996) 983–998.

[39] R.H. Logie, C. Marchetti, Visuo-spatial working memory: visual, spatial or central executive? in: R.H. Logie, M. Denis (Eds.), Mental Images and Human Cognition, Elsevier Science, 1991.

[40] J.M. Loomis, R.L. Klatzky, S.J. Lederman, Similarity of tactual and visual picture recognition with limited field of view, Perception 16 (1991) 106–120.

[41] P. Mahrer, C. Miles, Recognition memory for tactile sequences, Memory 10 (2002) 7–20.

[42] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.

[43] R. McKeon, The Basic Works of Aristotle, Random House, New York, 1941.

[44] M.A. Meredith, On the neuronal basis for multisensory convergence: a brief overview, Cog. Brain. Res. 14 (2002) 31–40.

[45] S. Millar, Memory in touch, Psicothema 11 (1999) 747–767.

[46] F.N. Newell, J.M. Findlay, The effect of depth rotation on object identification, Perception 26 (1997) 1231–1257.

[47] F.N. Newell, M.O. Ernst, B.S. Tjan, H.H. Bülthoff, Viewpoint dependence in visual and haptic object recognition, Psych. Sci. 12 (2001) 37–42.

[48] F.N. Newell, A.T. Woods, M. Mernagh, H.H. Bülthoff, Visual, haptic and cross-modal recognition of scenes, 2004, manuscript submitted.

[49] P. Patton, K. Belkacem-Boussaid, T.J. Anastasio, Multimodality in the superior colliculus: an information theoretic analysis, Cog. Brain. Res. 14 (2002) 10–19.

[50] V.S. Ramachandran, D. Rogers-Ramachandran, M. Stewart, Perceptual correlates of massive cortical reorganization, Science 258 (1992) 1159–1160.

[51] K. Rayner, A. Pollatsek, Eye-movements and scene perception, Canad. J. Psych. 46 (1992) 342–376.

[52] J.M. Reales, S. Ballesteros, Implicit and explicit memory for visual and haptic objects: cross-modal priming depends on structural descriptions, J. Exp. Psych. Learn. Mem. Cog. 25 (1999) 644–663.

[53] I. Rock, J. Victor, Vision and touch: an experimentally created conflict between the two senses, Science 143 (1964) 594–596.

[54] D.J. Simons, D.T. Levin, Change blindness, Trends Cognit. Sci. 1 (1997) 261–267.

[55] D.J. Simons, R.F. Wang, Perceiving real-world viewpoint changes, Psych. Sci. 9 (1998) 315–320.

[56] B.E. Stein, M.A. Meredith, The Merging of the Senses, MIT Press, Cambridge, MA, 1993.

[57] E.V. Sullivan, M.T. Turvey, Short-term retention of tactile stimulation, Quart. J. Exp. Psych. 24 (1972) 253–261.

[58] M.J. Tarr, H.H. Bülthoff, Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993), J. Exp. Psych. Hum. Percept. Perform. 21 (1995) 1494–1505.

[59] A.M. Treisman, Features and objects in visual processing, Sci. Amer. 255 (1986) 114–125.

[60] S. Ullman, High-level Vision: Object Recognition and Visual Cognition, MIT Press, Cambridge, MA, USA, 1996.

[61] A. Zangaladze, C.M. Epstein, S.T. Grafton, K. Sathian, Involvement of visual cortex in tactile discrimination of orientation, Nature 40 (1999) 587–590.

[62] Y.-D. Zhou, J.M. Fuster, Neuronal activity of somatosensory cortex in a cross-modal (visuo-haptic) memory task, Exp. Brain Res. 116 (1997) 551–555.