UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Psychology

Combining Cues and Recalibrating Priors for Accurate Perception

by

Iona Stephanie Kerrigan

Thesis for the degree of Doctor of Philosophy

September 2012

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
Psychology

Doctor of Philosophy

COMBINING CUES AND RECALIBRATING PRIORS FOR ACCURATE PERCEPTION

by Iona Stephanie Kerrigan

Our senses allow us to identify objects, materials and events in the world around us, enabling us to interact effectively with our surroundings. However, perception is inherently ambiguous, in that each set of sensory data could have resulted from an infinite number of world states. To find statistically optimal solutions the brain uses the available sensory data together with its prior knowledge or experience. In addition, when there are multiple cues available they may be: (i) combined to improve precision through noise reduction; and/or (ii) recalibrated to improve accuracy through bias reduction. This thesis investigates cue combination, learning and recalibration through a series of four studies, using the Bayesian framework to model cues and their interactions.

The first study finds that haptic cues to material properties are combined with visual cues to affect estimates of object gloss. It also investigates how the binocular disparity of specular highlights affects gloss estimates. This is extended in the second study, which finds that the human visual system does not employ a full geometric model of specular highlight disparity when making shape and gloss estimates. The third study replicates and extends previous findings, that auditory and visual cues to temporal events are optimally combined in adults, by demonstrating that children also optimally combine auditory and visual cues. Both adults' and children's bimodal percepts are shown to be well predicted by a 'coupling prior' model of optimal partial cue combination. The fourth study finds that the visual system can learn and invoke two context-specific priors for illumination direction, using haptic shape cues to provide calibratory feedback during training. It also demonstrates that colour can be learnt as a contextual cue.

The results of all these studies are considered in the context of existing work, and ideas for future research are discussed.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Iona Stephanie Kerrigan, declare that the thesis entitled *Combining Cues and Recalibrating Priors for Accurate Perception* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:.....

Acknowledgements

First and foremost I would like to acknowledge my supervisor, Wendy Adams. I, and this thesis, have benefited greatly from her expertise, advice and encouragement. I am also very grateful to Erich Graf for the guidance and additional perspective he has given. It has been a pleasure working with them both.

Thanks also to John Frisby, Jim Stone and John Porrill for encouraging me to pursue a PhD and equipping me with the skills to do so.

Thanks to Matt Jones for the development of experimental code used in Chapters 2 and 3. My thanks also go to Aaron Shuai Chang for data collection for the supplementary experiment in Chapter 3 and Fiona Berry, Nesta Caiger, Emma Ryan and Katie Hobbs for their help with data collection for Chapter 4. Thanks to Ben Backus for helpful scientific discussions regarding Chapter 5. Thanks to the Economic and Social Research Council for funding this thesis through a studentship. Thanks to Lisa Henly for her outstanding administrative support.

I am very grateful to David Knill and Robbie Jacobs for hosting me at the Department of Brain & Cognitive Sciences, University of Rochester, USA. They were very supportive and taught me much about using the Bayesian framework for computational modelling. I am also grateful to the wider graduate community in Rochester for making me so welcome and for teaching me to play softball.

I am very appreciative of the friends I have made through my PhD and in particular all the denizens of 3109, past and present, whose camaraderie and good humour have been invaluable; special thanks go to my fellow lab members, Katie Gray and Jennifer Josephs.

Thanks to my family for their support and for having always encouraged me to ask questions. Thanks also to Alastair Kerrigan, for his care in proof-reading, extraordinary patience and continual belief.

Chapter 1

Literature Review

1.1 The Problem of Visual Perception

To survive and thrive in the world it is necessary to perceive and identify objects, materials and events. This task, which appears entirely mundane, is actually an incredible feat. Things in the world may: be made of different materials; have different sizes, shapes and functions; and be located at different depths or travelling in different directions. To perceive the world around us the signals received by sensory receptors must be translated into a coherent and stable percept. The signals received by sensory receptors are ambiguous as to what they represent in the real world. This ambiguity can be seen by considering the case of visual perception; one of the tasks for the visual system is to translate the two dimensional (2-D) image on the retina into a three dimensional (3-D) percept. Translation from 2-D to 3-D is an example of an ill posed problem (Poggio & Torre, 1984), with an infinite number of solutions. A problem is well posed when its solution: exists; is unique; and depends continuously on the data (Marroquin, Mitter & Poggio, 1987) and ill posed when it fails to meet one or more of these criteria. To relate this to vision and state it another way, any image on the retina could have arisen as the result of an infinite number of scenes in the real world, all of which produce the same image. Another example of an ill-posed perceptual problem is in factoring a luminance image into surface colour and illumination. A change in image luminance could result from a change in surface reflectance or a change in the angle of incidence of the light source (for example due to object shape as seen in Figure 1.1) or any combination of these factors.

Despite the infinite number of interpretations available, our experience is usually of a stable 3-D visual world.[1] The ambiguity in the retinal image is due to a lack of

[1]There are instances when two or more interpretations can be perceived (although not simultaneously): these scenes are known as bistable or multistable and the perceived object 'flips' between the two (or more) interpretations (Figure 1.2). However, multistable stimuli normally only exist under unusual viewing conditions, for example in experimental situations, where there are a very limited number of

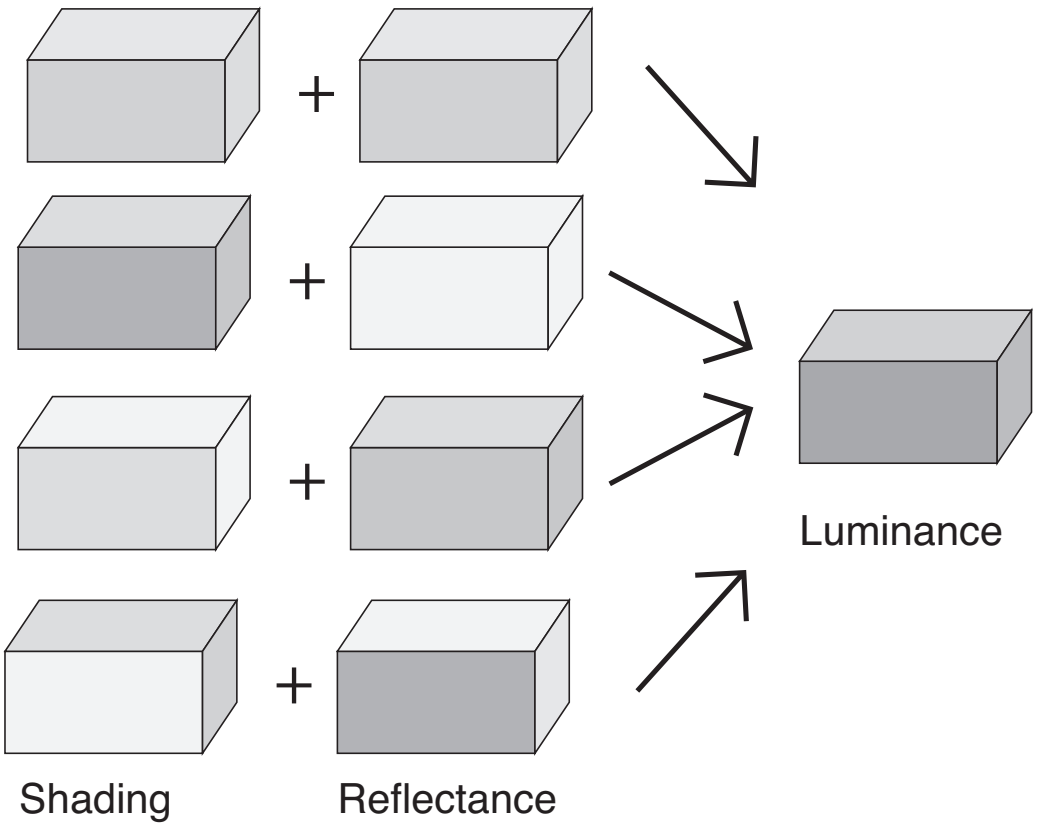

Figure 1.1: **Luminance as a result of shading or surface reflectance changes** To decompose the luminance image into its shading component and its reflectance component is a hard problem. The luminance image on the right could be formed from any of the shading and reflectance patterns on the left - or indeed any of an infinite number of other combinations of shading and reflectance. There is no way to determine directly, from a luminance measurement, what the reflectance properties of an object are since luminance changes may be as a result of a change in orientation with respect to the light source or of surface reflectance changes.

information: in the case of translation from 3-D to 2-D there is a loss of information and in the case of surface luminance there is a single measurement which is the result of several parameters (material properties, object shape, light intensity and lighting direction). One way to reduce the ambiguity in a scene, and to constrain the number of possible interpretations, is for perceptual systems to bring not only current measurements of scene variables but also their previous knowledge and experience to bear. In this review I will explore Bayesian approaches to the use of current sensory data and prior knowledge to acheive stable perception. Specifically, I will cover the following three ways in which perceptual systems can reduce uncertainty and error, using the Bayesian framework to:

- combine redundant signals, both within and across senses;

cues available. When viewing natural scenes people rarely experience any ambiguity since the visual system manages to solve the problem of perception sufficiently well under natural conditions.

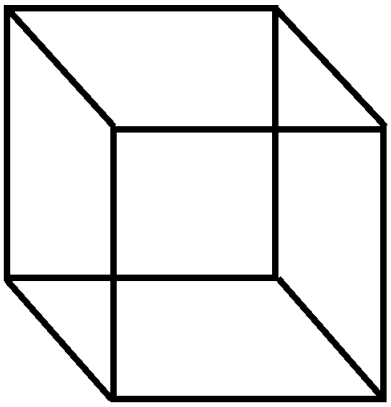

Figure 1.2: **Necker cube**
The Necker cube is a bistable stimulus in which the wire frame can be perceived as a 3-D cube with the front face either on the lower right or on the upper left of the image.

- incorporate prior knowledge to exploit the statistical regularities of the environment;
- and adapt to changes in statistical regularities (either by recalibration or learning of new relationships).

Although there are many ways to investigate this topic, I focus primarily on psychophysical experiments and computational modelling.

1.2 Visual Cues

1.2.1 What is a cue?

In a natural scene there are usually many sensory cues that are able to provide some information to constrain the number of possible interpretations. At the simplest level a perceptual cue is a sensed property that correlates with a world property. The types of world property that are useful for interaction relate to things like *what* something is (e.g., shape and/or material), *where* something is (e.g., depth) and *when* something happens (e.g., how many events and relative timing), in addition to the underlying causal relationships between objects and/or events. In the natural environment there are many cues, each potentially providing some useful information to enable inference of properties such as depth, shape and material.

A visual cue might be an image property (e.g., a distortion of a texture pattern) that over time has been observed to correlate with a scene property (e.g., depth or shape); on the other hand a visual cue could be a body property (e.g., accommodation) that correlates with a scene property (e.g., depth). The term 'cue' may thus refer to a wide

variety of different types of information, which can cause confusion. Knill, Kersten & Mamassian (1996a) set out useful criteria by which descriptors of visual (or any perceptual) information might be assessed; according to them descriptors of perceptual information should:

- be well-defined and logically consistent;
- relate the properties of the system's input to the environmental properties;
- state all prior assumptions about the environment and the system;
- be able to describe a wide range of types of information; and
- provide a language for specifying theories and generating testable hypotheses.

Visual cues, as described above, may not meet all of these criteria. They are ill-defined and may rely on unstated prior assumptions regarding the structure of the world. For example, texture is not a cue to shape or depth unless one makes assumptions about the texture's homogeneity across the surface. The Bayesian framework provides a principled method by which to describe perceptual cues in a manner which meets all of the above criteria. However, even within this framework it is somewhat arbitrary how different sensory information is segregated into cues by researchers and to some extent depends on conventional definitions. For example, there is no clear boundary between texture and linear perspective cues or between specular reflections and shading cues.

1.2.2 The Bayesian Framework

The Bayesian framework is a mathematical method for combining probability distributions to calculate other probability distributions that were previously unknown (Bayes, 1783). As it pertains to perception, Bayes' rule provides a formal method for combining current inputs to sensory systems with previous knowledge (Knill, Kersten & Yuille, 1996b; Maloney, 2002b; Mamassian, Landy & Maloney, 2002). The Bayesian model (shown in Equation 1.1) is composed of four probability distributions: posterior ($p(S|I)$); prior ($p(S)$); likelihood ($p(I|S)$); and evidence ($p(I)$).

$$p(S|I) = \frac{p(I|S)p(S)}{p(I)} \tag{1.1}$$

The evidence represents the probability of the sensory information being veridical or accurate: since the evidence is usually a constant, Bayes' rule can be simplified to Proportionality 1.2.

$$p(S|I) \propto p(I|S)p(S) \tag{1.2}$$

The likelihood probability distribution represents the probability of the current sensory information arising from any of an infinite number of world states (Maloney, 2002b). The likelihood function can also be thought of as the visual system's model of the statistical structure of the world, corrupted by noise. The peak of the likelihood function corresponds to the world state that is most likely to have caused the sensory data, assuming that all states are equally likely. The likelihood distribution can, therefore, be used directly to make estimates of world properties; selection of the world state most likely to have caused the sensory data is a decision rule known as Maximum Likelihood Estimation (MLE).

MLE does not, however, allow for the fact that different world states occur with very different probabilities. If the observer has knowledge of the relative prevalence of world states this can be used to bias estimates of world properties such as shape or depth. In the Bayesian model, this previous knowledge or experience of the world is represented by the prior probability distribution. A uniform prior probability distribution represents the case where all values of a scene property are equally probable (in the absence of any current sensory information); a non-uniform prior indicates that particular values are expected to occur more frequently than others. The prior encodes the relative probabilities of all possible states of a particular world property, independent from the current sensory data. In the example given in Figure 1.1, each of the shading and reflectance combinations shown are equally likely to have generated the luminance image. Using MLE it would be impossible to differentiate between the infinite number of such shading and reflectance combinations. However, incorporating the prior expectations that lighting is most often from above, and that reflectance is most often uniform across the surface of an object, the second option represents a much more probable world state and so should be preferred.

The combination of likelihood, prior and evidence gives the posterior distribution - the probability of a world property having particular values, given the current sensory information. A more refined estimate of a particular world property can then be made based on the position of the peak of the posterior probability distribution, a decision rule known as *Maximum a Posteriori* (MAP) estimation. This is likely to be a better estimate than with MLE as it takes more information into account (except where the prior is uniform, in which case the MAP and MLE estimators are equivalent).

A further refinement to this decision rule incorporates the concept of a utility or cost function, which is combined with the posterior to represent the expected gain or loss in acting upon a particular perceptual estimate (Maloney, 2002b). For example, the consequence of overestimating the width of a narrow ledge might be a dangerous fall; a useful cost function in this case might introduce a bias towards underestimating the width. Much more mundane scenarios can also benefit from cost functions: for example, when picking up a cup of tea there is some uncertainty in one's estimate of the cup's size; grasping with a grip aperture equal to that specified by the MAP

estimate may result in a spillage so some bias toward a larger size estimate may be introduced to minimise this possibility, as the probability of knocking over the cup is lower if the grip is too wide than if it is too narrow. In mathematical terms, a cost function represents the penalty incurred in interpreting a scene property as $\hat{x}$ when it is actually x (Berger, 1985); the utility function can then be defined as $U(\hat{x}, x) = -cost(\hat{x}, x)$ (i.e., the negative cost function). The decision rule in this case involves maximising expected gain by convolving the posterior probability distribution with the utility function and taking the mode of this new distribution as an estimate. Depending on the task, the costs and cost function might vary, and hence the decision rule might also vary.

The Bayesian framework provides a formal method for describing perceptual cues which meets all of the criteria set out by Knill et al. (1996a) (Section 1.2.1). In this framework, a cue can be thought of as the combination of likelihood and prior probability distributions. The likelihood distribution describes the relationship between system input and world properties; the prior distribution encodes assumptions about the environment or system. The Bayesian model is one of statistically optimal estimation, so provides a benchmark against which to assess human perception. Particularly important is that this approach provides a language for specifying theories and generating testable hypotheses.

1.2.3 Examples of Visual Cues

There are many different visual cues which can be used to help estimate a variety world properties. The following sections describe some examples of cues to shape and material, taking a Bayesian perspective of the information they provide. The cues I have selected are each used in the experimental chapters following this review.

1.2.3.1 Shading

Shading describes the variation in grey level across a surface due to the different orientations of parts of the surface relative to both the light source and observer (Mallot, 2000). Shading has long been known to be a cue to shape (Brewster, 1826) and has been used by many artists to create a sense of 3-D shape and depth in pictures. Luminance changes across an image might be the result of a change in the colour or material of an object (a reflectance change), however, they might also (or instead) be due to orientation changes of the surface with respect to the illuminant (Adelson & Pentland, 1996, see Figure 1.1 for an example of this). As mentioned in Section 1.1, separation of luminance changes into those due to surface reflectance and those due to surface orientation is an ill-posed problem.

For shading to be a cue to shape in a Bayesian sense it must be decomposed into a likelihood and a prior: the likelihood is the probability of the luminance variation given each combination of surface shape, lighting direction, lighting intensity and surface reflectance. Several assumptions must be represented as prior probability distributions. One such assumption is that there is only one light source in the scene (Ramachandran, 1988); Ramachandran (1988) showed that where stimuli had opposite shading patterns in the same scene, one of the shading patterns was interpreted as convex objects and the other shading pattern was interpreted as concave objects, implying an assumption of a single light source. Brewster (1826) found that object shape is interpreted as if the light source is above the object. More recent work has re-cast this assumption as the 'light-from-above' prior (Ramachandran, 1988; Kleffner & Ramachandran, 1992) or the 'light-from-above-left' prior (due to findings that in most observers there is a bias for them to interpret stimuli as though light comes from above left, Sun & Perona, 1998; Mamassian & Goutcher, 2001). Evidence that a single light prior is used across different tasks came when Adams (2007) found significant correlations between observers' light priors in a visual search task, shape judgements and reflectance judgements. However, although a single light prior may be used in different tasks, it is also flexible and can adapt to changing environmental statistics (Adams, Graf & Ernst, 2004; Adams, Kerrigan & Graf, 2010). Note that the light priors described here refer to probability distributions with 'above' or 'above-left' being the most probable lighting position, and 'below' or 'below-right' being the least probable (but still possible). When shading is interpreted in association with other shape cues it is therefore still possible to interpret the shading as consistent with light from other directions.

Another assumption that helps to interpret shading is that reflectance is constant across the surface of an object, such that gradual changes in luminance are likely to be due to shading (i.e., surface orientation changes) whilst abrupt changes in reflectance are usually due to changes in surface pigmentation (Adelson, 2000). Perhaps due to the number of assumptions that must be made, when other shape cues such as binocular disparity are present, shading proves to be a relatively weak cue: it has a relatively small effect on the final shape percept (Bülthoff & Mallot, 1988).

1.2.3.2 Texture

Texture refers to the spatial distribution of markings on a surface. The markings may exist as a result of the material properties of the surface itself (as in the variation of colour across a marble floor) or the arrangement of many similar objects on top of the surface (as in a gravelled path). Information regarding the shape and orientation of the surface can be determined from the distortion of such patterns due to perspective effects, however to do so requires certain assumptions about the regularity of the texture (Aloimonos, 1988).

There are three components of texture that can be used as cues to shape (Cutting & Millard, 1984; Blake, Bülthoff & Sheinberg, 1996; Knill, 1998): perspective scaling, in which the size of texture elements decreases with distance from the observer; perspective foreshortening, in which the aspect ratio of texture elements is compressed depending on the slant of the object; and the relative position, or density, of texture elements, which is affected by scene properties such as surface curvature. To make use of these effects of perspective projection, an observer must make assumptions about the statistical structure of the texture, namely in the homogeneity (even spacing) and isotropy (no directional bias) of texture elements (Blake et al., 1996). Deviations from homogeneity can then be used to make inferences of surface attitude (i.e., slant and tilt). Similarly, deviations from isotropy of texture elements can be used to infer foreshortening and surface orientation.

From a Bayesian point of view, each of the perspective effects described above may be modelled as a separate cue with corresponding likelihood and prior probability distributions. Knill (1998) presented a mathematical model for implementing these. In his model, the likelihoods represent the probability of the observed texture pattern observed given the slant and tilt of the object/surface. Similarly, the assumptions about homogeneity, isotropy, size and aspect ratio are modelled as prior probability distributions.

1.2.3.3 Specular Reflections

Specular reflections occur when light is reflected regularly, rather than diffusely, from the surface of an object. It is these reflections which give surfaces the impression of gloss or shininess; some materials (e.g., mercury mirrors) reflect all light regularly, whereas other materials (e.g., plastic surfaces) reflect only some of the light regularly and some diffusely. Specular reflections are not fixed to the surface of an object, as a texture pattern would be, rather they change shape and location with movement of the viewer, object or environment. To interpret the shape of an object by calculating the transformation that led to the distorted reflection would require the observer to have an accurate model of the surrounding world (Adelson, 2001). Given the multitude of different scenes in which a glossy object may be found, this would seem to be extraordinarily unlikely.

A different method to interpret shape from the distortions in specular reflections relies on the assumption that image statistics, such as amplitude spectra and distribution of orientations, are similar across different natural scenes (e.g., Torralba & Oliva, 2003; Field, 1987). When images are reflected, their statistics are distorted depending on the surface shape (Longuet-Higgins, 1960; Fleming, Torralba & Adelson, 2004). A completely flat mirrored surface will reflect exactly the statistics of the real world. However, if the surface is curved the image statistics will differ from natural scene

statistics as a function of that curvature in relation to viewing position, with the image compressed at certain points and stretched at others (Fleming et al., 2004). Fleming et al. (2004) predicted that if people possess a model of natural scene statistics then they should be able to estimate 3-D shape accurately, even without specific knowledge of what is in the surrounding scene. To test this prediction they generated irregular mirrored objects rendered within different real-world scenes. The objects were then removed from their context and participants adjusted surface normals to indicate their perception of the slant and tilt of the surface at different points on the object. They found that people could make shape judgments accurately based solely on specular reflections.

An additional prediction Fleming et al. (2004) made is that if the scene is structured, the distorted reflection should lead to characteristic orientation fields across the image. They suggest that people are able to extract information, from the orientation fields, about object shape in a similar manner to how they use texture. As has been mentioned already, for textured surfaces shape can be estimated from the pattern of compression and rarefaction of the texture across the image. On a mirrored surface, there is a different, but still systematic, relationship between the shape of the object and the distortion of natural scene statistics. Whereas texture is distorted by slant but not by surface curvature, distortions in specular highlights are caused by surface curvature but not by slant. Fleming et al. described how orientation fields are able to provide constraints on 3-D shape; however, there are still some ambiguities, such as the sign of the surface curvature (whether convex or concave). The local constraints provided by orientation fields can be disambiguated when in combination with other cues such as the bounding contour or binocular disparity.

As noted above, specular highlights and reflections are cues to material properties, as well as to shape, with objects tending to look glossier when they have specular highlights on the surface (Beck & Prazdny, 1981). Fleming & Bülthoff (2005) also found that adding specular reflections to objects increased the authenticity of the sense of translucency in materials. They suggest that this is because translucent materials are often glossy, for example plastic, jade or wax. There are various ways in which highlights might be used to assist in the evaluation of material properties. Motoyoshi, Nishida, Sharan & Adelson (2007) found that positive skew of the luminance histogram might be a relatively simple method of detecting whether a surface is glossy. The stimuli they used were photographs of painted stucco which varied in albedo and gloss depending on the type of paint used. In contrast, Anderson & Kim (2009) found that it was not the skewness of the luminance histogram that affected how glossy a surface looked but rather the position of the highlights relative to the shading gradient. Anderson & Kim (2009) split each image used by Motoyoshi et al. (2007) into its diffuse and specular components; they then rotated the specular highlight components and asked people to rate how glossy the surfaces looked. The further the highlights

were offset from the diffuse shading gradient, the less glossy the surfaces looked, implying that the highlights were interpreted as pigment changes rather than as specular highlights when they were not aligned with the shading. In later work, they found the same effect of reduced gloss when highlights were offset from the shading gradient but still consistent with the underlying shape of the object (Kim, Marlow & Anderson, 2011). Although specular highlights can be useful indicators of gloss and shape, in some situations they are uninformative and may even be problematic, for example when making lightness judgements; in these situations they can be discounted by the visual system (Todd, Norman & Mingolla, 2004).

1.2.3.4 Binocular Disparity

Binocular disparity is a visual cue which results from the fact that the eyes are at different positions on the head and therefore each sees the world from a different viewpoint (Wheatstone, 1838). It is different from the other cues discussed so far in that rather than arising from the retinal image directly, it results from the observer comparing the images from two slightly different view points. Absolute depth can be judged from disparity provided that one knows the distance between one's eyes. Since this distance increases during a person's lifetime, from about 4cm to about 6cm (MacLachlan & Howland, 2002), it suggests that binocular disparity must be recalibrated with reference to other cues over an extended period of time. Binocular disparity is a very useful and reliable cue at short range, however, the utility of binocular disparity as a cue to depth decreases with distance - precision decreases as the square of the distance to the object (Hillis, Watt, Landy & Banks, 2004). Reframing binocular disparity as a Bayesian cue the likelihood is the probability of the disparity field, given object shape or depth.

One of the difficulties for binocular disparity is knowing how to match up points in the left and right images: this is known as the 'correspondence problem'. Julesz (1960, 1964) showed that the visual system can solve the correspondence problem, even in the absence of other depth cues. He created random dot stereograms in which the left and right eye images are composed of almost identical configurations of randomly distributed dots on a plain background. Small differences in the positions of corresponding dots in the left and right eye images, that are undetectable when viewed monocularly, result in a variation in disparity across the image when the two are fused, such that depth and shape become apparent. Marr & Poggio (1976) used random dot stereograms to test their stereo correspondence algorithm which relied upon two assumptions: (i) 'uniqueness': that each point in the left eye's image may only correspond with, at most, one point from the right eye's image (and vice versa), since an object or feature can only exist at one location in space; and (ii) 'continuity': that disparity generally varies smoothly with relatively few discontinuities, which usually

indicate the edge or boundary of an object. Later models have refined and added to these assumptions, for example Pollard, Mayhew & Frisby (1985) used the same uniqueness assumption but added an 'epipolar constraint'. This is a geometric constraint resulting from the relative viewpoints of the two eyes: for a given point in the left eye's image, the corresponding point in the right eye's image must lie along a so-called 'epipolar' line. Pollard et al. also refined the surface continuity assumption to a local disparity gradient limit (Burt & Julesz, 1980) which is more robust to jagged surfaces.

The assumptions described above can help shape perception of matte objects, however, specular reflections are at a different depth to the surface that generates them and so can cause large depth discontinuities (and large disparity gradients) on a surface. Previously, these discontinuities were deemed to be unhelpful for shape perception (e.g., Oren & Nayar, 1996). More recently, however, it has been shown that since the disparity of a specular reflection varies from the disparity of the surface generating it in a regular and predictable manner, it can be used to suggest whether a surface is convex or concave, in addition to affecting gloss judgments of the surface (Blake & Bülthoff, 1990, 1991; Wendt, Faul & Mausfeld, 2008; Wendt, Faul, Ekroll & Mausfeld, 2010). If the surface is curved the disparity of the reflection will depend on the convexity or concavity of the surface. If a surface is convex the reflection will appear to be behind the surface (a virtual image); if, on the other hand, the surface is concave the reflection will appear to be in front of the surface (a real image, see Figure 1.3), except under very unusual viewing conditions. However, for complex glossy objects in complex light fields specular reflections generate disparity fields that may contain a wide range of disparities; the corresponding depth field may contain large discontinuities and be poorly defined or even undefined (Muryy, Fleming & Welchman, 2012).

1.3 Cue Combination

So far each cue has been described in isolation. However, we usually experience the world around us through a multitude of cues both from a single modality (e.g., vision) or across different sensory modalities. People seem to be able to usefully combine cues to perceive the world as a coherent, stable percept both within and across senses which provides a good reason to study how cues are used in combination. When there are multiple cues available people can use them in two different (although not necessarily mutually exclusive) ways: firstly they may combine the cues to improve precision, by reducing noise (e.g., Ernst & Banks, 2002); alternatively, they may be used to recalibrate one another to reduce longer-term bias and so increase accuracy in the future (e.g., Adams, Banks & van Ee, 2001). Precision relates to how noisy or uncertain a cue is; if there is little noise the cue estimate is reliable and has low uncertainty: the cue estimate is precise. A cue estimate may be very precise but

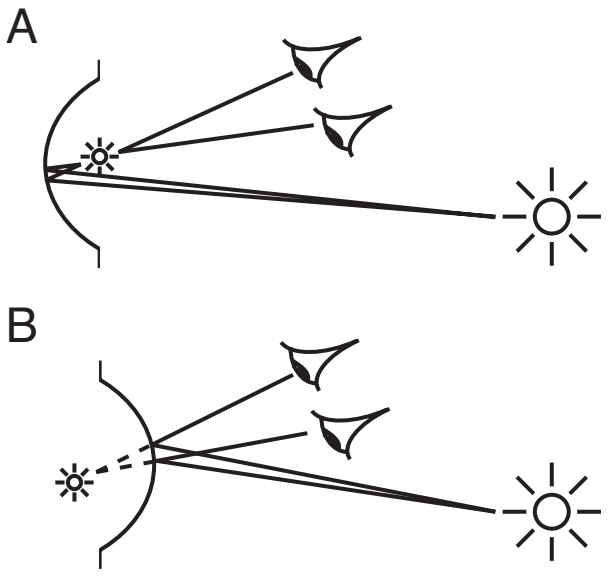

Figure 1.3: **Binocular geometry of specular highlight disparity**
The binocular geometry of specular highlight disparity on concave and convex surfaces, adapted and reprinted by permission from Macmillan Publishers Ltd: Nature (Blake & Bülthoff, 1990), copyright 1990. The highlight appears in front of a concave surface as a real image but behind a convex surface as a virtual image.

biased; accuracy relates to how biased a cue estimate is, that is how closely aligned the estimate is to the real world value. In this section I will review literature relating to noise reduction through cue integration (increasing precision); bias reduction through recalibration (increasing accuracy) is considered in Section 1.6.1.

1.3.1 Models of Cue Combination

There are two extreme views regarding the modularity of cue combination: if a system is modular then each 'module', in this case each cue, is independent of other modules. If cues are modular then each may be used independently to estimate a property, such as depth, before they are combined linearly: this is known as weakly coupled data fusion (or weak fusion; Clark & Yuille, 1990). An example would be that an estimate of shape was made from each of texture, binocular disparity and shading before being combined into a single estimate of shape. The alternative view of modularity in cue combination is known as strongly coupled data fusion (or strong fusion; Clark & Yuille, 1990). In strong fusion the output of sensory processing modules are able to affect the progress of other modules by interacting and altering constraints. Additionally, these outputs may be combined using any combination rule. In this approach cues are not

evaluated separately before combination but may influence one another throughout the combination process. Mathematically, the differences between these two approaches can be seen by comparing Equation 1.3 (weak fusion, the cue estimate ($\hat{S}$) equals the linear weighted average of modular cue estimates) and Equation 1.4 (strong fusion, the cue estimate ($\hat{S}$) is a function of all the different available 'cues' or information using an arbitrarily complex combination rule). Between these two extreme views there is a model known as modified weak fusion (Landy, Maloney, Johnston & Young, 1995). This is essentially a modular viewpoint but avoids some of the disadvantages of the weak fusion account. Each of these three models can be instantiated using the Bayesian framework; each model is described below and their relative strengths and weaknesses discussed.

$$\hat{S} = \frac{\sum_{i=1}^{n_{cues}} w_i f(cue_i)}{\sum_{i=1}^{n_{cues}} w_i} \tag{1.3}$$

$$\hat{S} = f(cue_1, cue_2, cue_3 ... cue_n) \tag{1.4}$$

1.3.1.1 Strong fusion

Strong fusion can be thought of as an extension to the single-cue Bayesian model, in which a single likelihood is used to encode the probability of the entire image given the range of possible world states. Strong fusion models assume that cues are not independently processed before being combined. Within this framework the separation of cues, such as disparity, shading and texture, is meaningless since cues are considered to be artificial constructs. The single likelihood thus encapsulates all cues within the scene, rather than an individual cue. Nakayama & Shimojo (1992) developed a simple strong fusion model that selects as its interpretation the scene that is most likely to have generated the image, i.e., it maximises the likelihood of the image given the scene. They make use of a generic viewpoint assumption, as one would typically have when mobile, but otherwise omit any prior knowledge that could be incorrect. This model is essentially the same as the MLE approach to single cue estimation (see Section 1.2.2).

Yuille & Bülthoff (1996) describe a similar model for shading and texture integration in which the luminance and hue of an object cannot be factored into the two components (shading an texture) without having first determined the object shape. To calculate shape from shading requires an assumption of constant albedo across the surface, which texture always violates. They argue that this inseparability requires a joint likelihood function and so a strong fusion account is required to model the combination. Bülthoff & Mallot (1990) showed that observers underestimated the elongation of an ellipsoid

when using either shading or texture in isolation, but were less biased when both were present. A weak model of fusion would predict lower response variance but similar bias in the multi-cue condition; a strong model of fusion, in which cues are able to interact in a non-linear fashion, seems better able to account for their data. An alternative explanation is that observers were using an unmodelled prior, e.g., for regular (in this case, spherical) shapes, which had more influence on the final shape estimate when only a single cue was available. When both cues were available, the relative influence of the prior may have reduced resulting in less biased estimation of elongation.

The strong fusion model also allows for the inclusion of prior information and the use of a MAP estimator; in such cases there is a single prior probability distribution representing the relative probabilities of each possible world state. Nakayama & Shimojo (1992) acknowledged that this would be possible but argued that it was not appropriate for the perception of visual surfaces due to the inherent difficulty in estimating the prior. For example, the prevalence of particular world properties varies in different environmental contexts; the prior would therefore need to vary also, making it unmanageably complex. Alternatively, multiple priors would need to be learned for the various contexts; I will revisit the issue of learning multiple priors for different environmental contexts in Section 1.6.2 and Chapter 5.

It is difficult, from a modelling perspective, to empirically test the strong fusion account since the cues are not separable from one another in a controlled manner. The interactions between cues may also be arbitrarily complex, so a model describing the combination of two cues would not be able to predict performance when a third is added. However, the strong fusion account has several advantages over more modular approaches: in particular, it allows cues which provide information about qualitatively different world properties (for example, relative vs. absolute depth) to constrain each other. Chapter 2 describes a study involving combination of two cues (visual and haptic) to estimate a purely visual world property (gloss); this could be considered as strong fusion due to the fundamentally different types of information, however, see also Section 1.3.1.3 for an alternative approach.

1.3.1.2 Weak fusion

At the other end of the spectrum, the weak fusion model treats each cue as a completely independent module which, in Bayesian terms, has its own likelihood and prior probability distributions. In the simplest form of weak fusion, the final estimate of world state is calculated by generating an individual MAP estimate from each cue and combining these through linear averaging (e.g., Clark & Yuille (1990)). As there is redundant information across the different cues, this linear combination will result in a reduction in the variance of estimates. However, an additional consequence of this linear combination is that the final estimate will always lie between the two individual

cue estimates. This form of the weak fusion account is therefore unable to explain the results of Bülthoff & Mallot (1990) as described in Section 1.3.1.1.

A more general approach to modular cue combination can be achieved by calculating a joint posterior probability distribution equal to the product of all of the likelihoods and priors of the cues being modelled (Equation 1.5).

$$p_{1,2}(S|I) \propto p_1(I|S)p_1(S)p_2(I|S)p_2(S) \tag{1.5}$$

In the case where the individual estimates are similar and the probability distributions are all Gaussian, Yuille & Bülthoff (1996) showed that maximising the joint posterior is equivalent to the simple case of linear averaging described above, where the individual estimates are weighted according to their reliability (inverse variance). These assumptions were used by Hillis et al. (2004) to investigate the combination of texture and disparity cues to slant. However, the form given in Equation 1.5 is more flexible as it does not require that the estimates are similar, or place any constraints on the parametric form of the likelihoods or priors.

Adams & Mamassian (2004) showed that the simple linear averaging approach could not account for the combination of texture and disparity cues to shape. Observers estimated shape from texture for ambiguous stimuli which were consistent with either a convex or concave surface: they consistently reported seeing convex shapes. This is consistent with a prior for convexity, and a bimodal likelihood distribution in which the convex and concave interpretations are equally likely to have caused the same image. The stimuli were subsequently disambiguated by adding binocular disparity. In the case of a concave disparity-defined shape, a cue combination rule based on weighted averaging would predict that as texture-defined curvature increased, the overall shape estimate would reduce in concavity (i.e., the surface would look flatter) since the texture estimate would be increasingly convex. Adams & Mamassian instead found that such stimuli appeared more concave as the texture-defined curvature increased: the disparity was disambiguating the bimodal texture cue. Equation 1.5 is still applicable in this case, but is no longer consistent with a linear weighted average as the texture likelihood function is not Gaussian in form.

From an empirical perspective the advantages of weak fusion are that it is modular and the combination rule is simple: if weak fusion occurs then it is meaningful to study each cue in isolation. The primary problem for the visual system with the weak fusion account is that different cues contain qualitatively different information. For example, when estimating depth information one cue may give absolute depth (e.g., binocular disparity) information whilst another may give only relative depth (e.g., texture). This qualitative difference means that it is not always meaningful to average across estimates of different cues (Landy et al., 1995).

1.3.1.3 Modified weak fusion

In response to accounts of strong and weak fusion Landy et al. (1995) put forward the idea of modified weak fusion (MWF). They consider strong and weak fusion to be at opposite ends of a spectrum. Modified weak fusion retains the modelling advantages of weak fusion, such as modularity and a linear combination rule. It also overcomes some of the problems inherent for the visual system in weak fusion by allowing some interactions between qualitatively different cues as required to transform all the cue estimates into the same domain. They call this cue 'promotion' - a cue is promoted when it has missing parameters filled in by other cues in order that the information is of the same type, for example a relative cue to size or distance, such as texture, might be promoted, that is have extra constraints added from other available cues, to an absolute metric cue. This appears to be an example of a strong fusion type interaction; however, in their model, Landy et al. (1995) allow this kind of interaction to occur only for the purposes of promotion. Evidence of 'strong' interactions beyond these circumstances would falsify the MWF account.

Adams & Mamassian (2004) argued that explicit cue promotion is not necessary if additional prior information is incorporated into the model. However, observers in their study may have used knowledge of the depth of the screen itself (e.g., through accommodation cues) to promote texture from a relative to an absolute cue to stimulus depth before combination with binocular disparity (another absolute depth cue). Texture may in effect already have been promoted before observers' estimates of depth were reported.

When cues give rise to the same types of information, weak fusion is indistinguishable from modified weak fusion as there is no need for cue promotion. Several studies (e.g., Adams & Mamassian, 2004; Maloney, 2002a) simply assume that all cues produce estimates in the same units (i.e., that cues have already been promoted), whereas others deliberately choose to investigate cues to the same information type (e.g., Young, Landy & Maloney, 1993). Hillis et al. (2004) found evidence that disparity gradient had been promoted to units of surface slant, so that it could be treated as a direct cue to slant in their combination model. Whilst theoretically more complete than the weak fusion account, the cue promotion component of MWF is often not explicitly modelled: the majority of studies of cue combination acknowledge the need for cue promotion in general but appear to implement a weak fusion model anyway.

1.3.2 Benefits of Cue Combination

In many circumstances the more constrained variant of weak fusion, in which cues provide the same type of information, give similar estimates and have Gaussian likelihood and prior probability distributions, is applicable. Where cues can be

assumed to have been promoted already (e.g., Maloney, 2002a) this approach is also applicable. Note that many of the benefits described below also apply to more complex variants of weak and modified weak fusion, where the constraints listed above are not met.

Estimates of world properties made from any single cue are prone to error and may not exactly represent the real value of the property in the world. There are two potential sources of error in the estimation of a world property from a cue: bias and noise. Bias is a systematic error in which the estimate is consistently incorrect due to poor calibration of a perceptual cue with respect to the estimated world property. If estimates from two cues are linearly combined, any bias may average out if the bias is in different directions. However, it is also possible that several cues may be biased in the same direction in which case combining cues would offer little or no benefit in bias reduction.

Noise, on the other hand, results from random measurement error and may vary from one moment to the next (Hillis, Ernst, Banks & Landy, 2002). The variance of a perceptual cue's posterior probability distribution is a measure of the noise associated with that cue; a posterior with high variance will give rise to a noisy estimate of the underlying world property value. A cue's reliability can therefore be defined as the inverse of its variance. Combination of multiple noisy cues will result in a more precise and reliable estimate if certain assumptions are met. In particular, if the posterior probability distributions from which the individual estimates would be drawn are independent of one another, and are Gaussian in form, then the product of the two distributions will have a variance less than or equal to that of the lower of the two individual variances. In other words, the reliability of the combined estimate is at least as high as (and usually higher than) that of an individual estimate from the more reliable cue. The combined estimate (the peak of the combined posterior probability distribution) will lie between the two individual estimates, and nearer the more reliable of the two (see Figure 1.4). When combining uncorrelated Gaussian distributions in this way, the result is mathematically equivalent to calculating the linear weighted average of the two individual estimates, with each component weighted in proportion to its reliability as defined above (Ernst & Banks, 2002). More formally, this linear combination rule can be described as in Equation 1.6:

$$S(d,t) = w_d S_d(d) + w_t S_t(t) \tag{1.6}$$

Where $S(d,t)$ is the combined estimate from two cues, for instance, shape from disparity and texture, $S_d(d)$ is the estimate from disparity, $S_t(t)$ is the estimate from texture and these estimates are weighted by w_d and w_t respectively. The values of w_d and w_t are based on the estimated reliability of each cue.

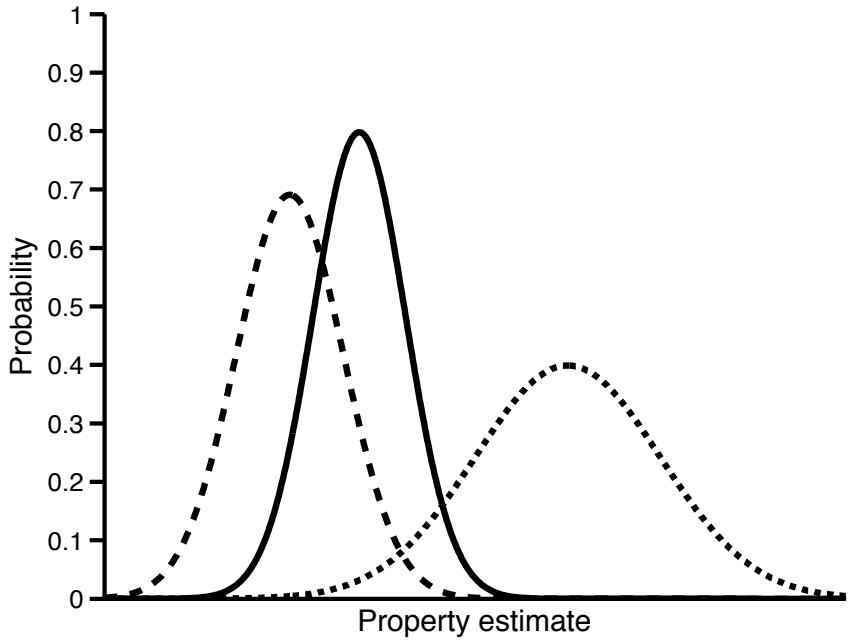

Figure 1.4: **Combination of cue estimates**
Illustration of how combining two cue estimates, differing in reliability, results in less uncertainty in the combined estimate. This graph shows two individual posterior distributions (dashed lines) from two cues; each is Gaussian, and the two are uncorrelated. The product of the two distributions is also shown (solid line); its peak (i.e., the estimate) lies between the two distributions, biased towards the more reliable cue, and the joint estimate has a smaller variance than either. This is equivalent to reliability based cue weighting and is a consequence of multiplying together two Gaussians with uncorrelated noise.

If $\hat{S}_d$ is the optimal disparity estimate of shape, that is, the estimate that maximises the probability of shape value given disparity cue $p(S|d)$, and $\hat{S}_t$ is the optimal texture estimate of shape, that similarly maximises $p(S|t)$; then, $\hat{S}$ is the optimal shape estimate based on disparity and texture - that is the shape S that maximises $p(S|d,t)$. Yuille & Bülthoff (1996) showed that:

$$\hat{S} = w_d \hat{S}_d + w_t \hat{S}_t \tag{1.7}$$

where,

$$w_d = \frac{\frac{1}{\sigma_d^2}}{\frac{1}{\sigma_d^2} + \frac{1}{\sigma_t^2}} \text{ and } w_t = \frac{\frac{1}{\sigma_t^2}}{\frac{1}{\sigma_t^2} + \frac{1}{\sigma_d^2}} \tag{1.8}$$

and σ_d^2 and σ_t^2 are the variances of the distributions $p(S|d)$ and $p(S|t)$ respectively. By choosing this way to specify the weights it means they are directly related to the reliability of the cues themselves. The other advantage of this method is that weights

are non-negative and sum to one. By weighting the cues in this way it ensures that the variance in the final estimate is less than or equal to the variance in the most reliable estimate; the method by which the combined variance (σ^2_{dt}) can be calculated is shown in Equation 1.9.

$$\sigma^2_{dt} = \frac{\sigma^2_d \sigma^2_t}{\sigma^2_d + \sigma^2_t} \tag{1.9}$$

Combining estimates using the method described here can be thought of as 'statistically optimal' cue integration since it provides a minimal variance unbiased estimator of a world property (Clark & Yuille, 1990; Landy et al., 1995). It is also commonly referred to as an 'ideal observer' model of cue integration (e.g., Landy et al., 1995; Hillis et al., 2004; Ernst & Bülthoff, 2004)

1.4 Examples of Cue Integration

Two benefits of cue integration have been outlined: reduction in systematic error (bias) and random error (variance). The issue of bias reduction and (re)calibration will be returned to in Section 1.6.1. The following sections describe examples of cue integration both within a single modality and across different sensory modalities.

1.4.1 Within-modality

Several studies have shown that visual cues are combined in a way consistent with optimal cue integration. Landy & Kojima (2001) used two different texture cues, spatial frequency and orientation of texture elements. Two textures were presented in each stimulus such that there was a boundary part way across the stimulus. Observers indicated in which of two stimuli (presented simultaneously, one above the other) the boundary appeared further to the left. In common with later cue integration studies, cues were presented either individually or in combination. The single-cue trials were used to calculate discrimination thresholds: a measure of the variance of estimates associated with each cue in isolation. The individual cue estimates and their respective variances were then used to generate predictions for combined estimates in the case where the two cues were presented together, and in particular when they were in conflict (i.e., the spatial frequency and texture element orientation cues specified different locations for the boundary in a single stimulus). This cue conflict paradigm is commonly used to study cue integration as it allows the relative contributions of each cue to the combined estimate to be determined. If two cues both specify the same value of a world property, then the individual and combined mean estimates will all be approximately equal (assuming unbiased estimators and negligible effects of priors or

other, unmodelled cues) regardless of the weights used in that combination. However, if the two cues are contrived so as to specify different values, then the combined estimate will lie somewhere between the two individual estimates, at a value dependent on the relative weights given to those individual estimates. Optimal cue integration of this sort may be seen in Figure 1.4: the location of the texture boundary based on spatial frequency could be represented by the large dashed line on the left; the location of the texture boundary based on texture element orientation could be represented by the dotted line on the right. The final estimate (solid line) is the weighted average of the two cue estimates where those weights are inversely proportional to the individual cue estimate variances. Landy & Kojima (2001) found that this optimal integration model predicted their cue conflict data better than a 'switching' model, in which responses were drawn variously from the two single-cue distributions with a probability dependent on their relative reliabilities. They also compared a number of more complex models, some of which fitted the data better than the linear weighting model; however, they acknowledge that this is unsurprising given the increased number of free parameters available to optimise the fits of such models.

A similar approach was used by Knill & Saunders (2003) to study the combination of texture and disparity cues to surface slant. In addition to using the cue conflict paradigm as described above, Knill & Saunders varied the reliability of the texture cues by altering the slant of a surface (disparity cues maintained approximately the same reliability across different slants). The reliabilities of texture and disparity cues to slant were measured using discrimination thresholds for each cue, for each participant. The variation of texture cue reliability allowed them to test the hypothesis that weights are not fixed, but rather are proportional to the current cue reliability. They found that when presented with both cues together, observers applied different weights to each cue according to the reliability of the current stimulus, and combined the cues in the statistically optimal fashion described in Section 1.3.2. Hillis et al. (2004) reported similar findings when varying the reliability of both texture and disparity cues. Texture reliability was again manipulated by varying surface slant; in addition, the reliability of disparity was manipulated by varying the viewing distance. In both cases, observers dynamically adjusted the weight given to each estimate, suggesting that they were able to estimate the reliability of each cue on a trial-by-trial basis. An alternative possibility is that cue weights are not explicitly calculated but rather, probability distributions are represented neurally and are directly combined with one another (Pouget, Dayan & Zemel, 2003; Knill & Pouget, 2004).

Visual cues are always spatially and temporally co-located because observers have only a small portion of the visual scene in focus at one time. In most real world situations, co-located signals have a single underlying cause and any (small) differences are due to bias and noise. However, in experimental situations it is possible to introduce unusually large conflicts between cues so that they are cues to different underlying

values of the property (as described above). So far only the benefits of cue combination have been discussed; the fusion of two cues, though, also comes with a potential cost: the observer may lose access to individual cue estimates as described below. If two cues (S_1 and S_2) lead to a combined estimate ($\hat{S}_{1,2}$) the weak fusion combination rule is:

$$\hat{S}_{1,2} = w_1\hat{S}_1 + w_2\hat{S}_2 \tag{1.10}$$

Hillis et al. (2002) showed that it is possible to make changes to the values of S_1 and S_2 such that $\hat{S}_{1,2}$ is constant; to keep $\hat{S}_{1,2}$ constant, for a change in S_1 of ΔS_1:

$$\Delta S_2 = -\frac{w_1}{w_2}\Delta S_1 \tag{1.11}$$

If only the joint estimate ($\hat{S}_{1,2}$) is available to the visual system, the observer would be unable to discriminate between stimuli that are adjusted in a way consistent with Equation 1.11, even though she would be able to discriminate were either cue presented individually. Stimuli that are physically different from one other that cannot be distinguished perceptually are known as 'metamers' (Hillis et al., 2002). Hillis et al. took advantage of this fact to examine whether cue estimates to slant were completely fused or whether observers retained access to individual cue estimates (disparity and texture). Using the single cue discrimination thresholds and variances they predicted which stimuli would be discriminable if only the joint estimate were available and which would be discriminable if cues were not fused. They found that visual cues were fused such that there was no access to the individual cue estimates. This finding raises the question as to how cues might recalibrate one another if they are completely fused: without separate estimates there would be no error signal to drive recalibration.

Specular reflections and highlights provide cues to both shape and material as described in Section 1.2.3.3 and Section 1.2.3.4. Cue combination for material properties has not been widely studied (Adelson, 2001), particularly from an ideal observer perspective. Highlight disparity is known to be a cue to gloss (Blake & Bülthoff, 1990, 1991; Wendt et al., 2008, 2010) and can also be used as a cue to shape (Blake & Bülthoff, 1990, 1991). Blake & Bülthoff found that highlight disparity could drive shape interpretation of ambiguous shapes (convex vs. concave). However, although observers adjusted highlight disparity appropriately to increase gloss ratings for convex shapes, highlight disparity was not adjusted in line with physical geometry for concave shapes. They suggested that this was due to uncertainty in shape due to inadequate rendering methods and interpret their findings as though observers generate a single combination of shape and gloss estimates that is consistent with their estimate of disparity. Chapter 3 tests whether shape uncertainty can account for their findings by manipulating the reliability of shape cues and highlight disparity and asking observers for both an estimate of shape and an estimate of gloss.

1.4.2 Cross-modality

Just as there are many visual cues to world properties, in normal experience we often also have access to cues from each of the other sensory modalities. Our sensory experience is that of a coherent world, in which information across all our senses is integrated. Early studies of cross-modal cue integration found evidence to suggest that, rather than being combined, vision tended to dominate haptic information in shape estimates (e.g., Rock & Victor, 1964). Rock & Victor presented an object to observers; the object was visually distorted, observers also felt the object and were asked to estimate its shape, either by drawing it or matching it to another object. They found that how the object looked was more important when making a decision as to which other object matched it. This type of finding was initially termed visual capture (Pick, Warren & Hay, 1969) in a demonstration of the ventriloquist effect, which showed the dominance of visual cues over auditory cues. The ventriloquist effect is the phenomenon by which a ventriloquist makes it appear as though her voice comes from her puppet rather than from herself. The perceptual system assumes a single source due to the temporal synchrony between the visual motion of the puppet and the words spoken by the ventriloquist. Estimates of spatial location from audition are worse than those from vision and so the visual estimate of location 'captures' the auditory estimate of location. Visual capture can be thought of as the idea that visual cues are in some sense 'special' and will always dominate cues from other modalities when in conflict. Subsequently, this view has been proved wrong by demonstrating capture by other senses, for example auditory capture of vision for temporal rather than spatial estimates (Morein-Zamir, Soto-Faraco & Kingstone, 2003).

An alternative interpretation of visual capture in the ventriloquist effect is offered by Alais & Burr (2004). They demonstrated that the ventriloquist effect may be due to nearly optimal bimodal integration of the auditory and visual cues. They conducted an experiment in which participants were required to indicate where light 'blobs' or sound 'clicks' were in space. The light and sound were presented either separately (unimodally) or together (bimodally). Two stimuli were presented; the task was to state which of two presentations was further left. In just the same way as for within-modality cue combination studies, they estimated the variance associated with the estimate from each modality in the unimodal task. Subsequently, localisation for bimodal stimuli was measured. In one presentation auditory and visual cues were consistent with each other; in the second presentation the auditory and visual stimuli were displaced from each other so that there was a conflict in the position of the two stimuli. By comparing which of the two presentations appeared to be further left, it was possible to estimate the weight attached to each stimulus. It was found that visual cues and auditory cues are combined such that the more reliable cue dominates the less reliable cue. When visual cues were reliable, vision dominated, however, when visual cues were blurred such that they were a less reliable cue to location, audition

dominated. When the two cues were comparable in terms of reliability the cues were weighted such that neither sense dominated: equal weighted averaging occurred. In this study people were explicitly asked 'to envisage each presentation as a single event'. It is possible that this instruction increased the integration of the two cues whereas under natural viewing conditions there would be less integration. This interpretation of a dynamic re-weighting of cues is able to explain why sometimes one sense 'captures' another and sometimes the pattern reverses. Capture can thus be thought of as an extreme case of cue combination resulting from large differences in the reliabilities of the cues used.

Ernst & Banks (2002) found evidence that people also use this statistically optimal strategy when integrating visual and haptic cues to size. They estimated the reliability (inverse variance) of within-modality judgements using discrimination thresholds as an index. Performance in bimodal trials was predicted from the reliability of each unimodal judgement using the optimal cue combination model described in Section 1.3.2. Observers looked at, or felt, a raised ridge and made a judgement of its height compared with another ridge that they perceived in the same modality. The reliability of the visual cue was varied by altering the amount of visual noise in the scene. As the noise in the visual stimulus increased, its reliability decreased. In bimodal trials observers both saw and felt the stimulus. The same paradigm was used as in the study by Alais & Burr (2004) described above: in one presentation the two cues specified equal heights; in the other presentation of each trial the two cues each specified different heights. Which stimulus was judged to be taller depended on the weight given to each size cue. The observed and predicted discrimination thresholds for bimodal trials were compared and found not to vary significantly from one another, again suggesting optimal reliability based cue weighting.

The similarities between theories of cross-modal integration and integration of cues within a modality can be seen in that, in both, cues are weighted based on their reliabilities. Exactly the same mathematical models as posited for intra-modality cue combination can be applied to cross-modality combination. The same benefit of noise reduction applies to cue combination across modalities as within a single modality; however, as mentioned before this comes at a potential cost in that observers may lose access to the individual cue estimates. As described in Section 1.4.1, Hillis et al. (2002) found that people have no access to individual cue estimates within a single modality (vision). Whereas for visual-visual cues, mandatory fusion is not particularly costly, as all cue estimates are derived from the same retinal image and correspond to the same spatial and temporal location, this is often not the case for cues across different modalities. For example, we commonly interact haptically with one object whilst looking at another; to combine visual and haptic cues to object shape in this scenario would be detrimental. Hillis et al. found that although visual and haptic cues to

object size could be combined to reduce noise, observers did not lose access to individual cue estimates from each modality.

To decide whether to combine cues, perceptual systems must assess whether two cues were caused by the same object or event in the world. Within a modality it is usually apparent whether two cues have a common cause since they are spatially and temporally synchronous as measured by a single set of sensory apparatus. For auditory and visual stimuli, cues are more likely to be combined across modalities when they originate in the same place and at the same time than when they do not (Wallace, Roberson, Hairston, Stein, Vaughan & Schrillo, 2004). Similarly, visual-haptic cue integration decreases with increasing spatial separation when making object size judgements (Gepshtein, Burge, Ernst & Banks, 2005). Helbig & Ernst (2007a) created visual-haptic stimuli with conflicting cues to object size by distorting the visual image with a cylindrical lens; they also introduced a spatial separation between the visual and haptic objects using a mirror. Observers continued to integrate cues when making size judgements across the two modalities if they could see their own hand touching the displaced visual object. This suggests that prior knowledge that the visual and haptic objects are one and the same reduces sensitivity to spatial separation. A common scenario in which visual and haptic stimuli originate from the same object but are spatially separated is in tool use; Takahashi, Diedrichsen & Watt (2009) found that when a tool was used, the degree to which object size cues were integrated across modalities depended on the spatial separation of the tip of the tool and the visual object, rather than the separation of the hand holding the tool and the visual object. These studies suggest a flexible and somewhat complex approach to causal inference.

In addition to spatial separation of cues, cues may also be temporally offset from one another. A similar process is necessary in this case to decide whether the two cues are due to the same event or different events. This can be seen in the 'bounce/stream' illusion (Sekuler, Sekuler & Lau, 1997). In this illusion two discs move towards each other, meet and then move away from each other in a manner consistent with either bouncing off each other or streaming past each other. The proportion of trials for which the percept is of the discs bouncing off each other can be increased by the addition of an auditory stimulus played at or near the time of coincidence of the two discs. The closer the alignment between the auditory stimulus and the discs coinciding, the higher the proportion of 'bouncing' percepts. This result implies that temporal alignment is used as a cue to causal relationships between stimuli in different modalities. A further example of this can be seen in the flash/beep illusion (Shams, Kamitani & Shimojo, 2000): a single flash presented with two beeps tends to appear as two flashes. Similarly, if two flashes are presented with a single beep they tend to appear as a single flash (Andersen, Tiippana & Sams, 2004). Shams, Ma & Beierholm (2005b) modelled this as partial cue integration, where the amount or strength of integration is modelled as a joint prior on how likely it is that the audio and visual

events have a common cause. Bresciani, Dammeier & Ernst (2006) demonstrated a similar illusion using visual-haptic stimuli: the beeps were replaced with taps. They also modelled it as partial integration but using a 'coupling prior' to represent the influence of one modality on the other (this model is described in more detail in Chapter 4). Both Shams et al. (2005b) and Bresciani et al. (2006) extend the ideal observer model of cue combination to take into consideration the circumstances in which observers should not combine cues as well as when they should. Partial cue integration using a coupling prior will be considered again in Chapter 4.

The combinations described here are all for stimuli in which the world property value can be estimated from either modality. For example, slant, size and number of events are all things which have meaning in both modalities. There are also interactions between sensory cues which are less clear cut in terms of combining estimates. For example, the perceived sweetness and saltiness of food is affected by the level of background noise (Woods, Poliakoff, Lloyd, Kuenzel, Hodson, Gonda, Batchelor, Dijksterhuis & Thomas, 2011): presenting a sound and asking how sweet or salty it was would have no meaning. This type of interaction is somewhat unexpected but highly consistent between people; such interactions have been termed 'cross-modal correspondences' (e.g., Spence, 2011). Cross-modal correspondences have been found for many different sensory combinations, for example, odour with colour (Demattè, Sanabria & Spence, 2006) and odour with touch (softness) (Demattè, Sanabria, Sugarman & Spence, 2006), for further examples see Spence (2011). Cross-modal interactions or correspondences are interesting, particularly where a cue to a property only has meaning in one modality but has correlates in another. This could be the case in material perception where reflectance properties might have auditory or haptic correlates. Chapter 2 investigates whether the haptic cues of friction and compliance affect the visual perception of gloss.

1.5 Development of Cue Integration

The preceding sections have made clear that adults are able to combine perceptual cues in an optimal manner both within and across modalities. However, there has been limited research into the development of this ability. Nardini, Bedford & Mareschal (2010) compared the performance of adults with that of 6 to 12 year old children in a slant discrimination task, using texture and disparity as cues to slant. They measured discrimination thresholds for each cue individually, and compared these with discrimination thresholds when both cues were present. If both cues are integrated optimally, discrimination thresholds would be expected to decrease in magnitude, such that smaller differences in slant could be detected for the multi-cue stimuli compared with either cue in isolation. Nardini et al. found this to be the case for both adults and 12 year old children, however this was not true for 6, 8 or 10 year old children.

They did find a significant difference in the discrimination performance of 8 year old children between the disparity-only and combined conditions (disparity being the less reliable of the two cues when viewed in isolation), suggesting some cue combination although non-optimal. This comparison was not significant in 6 and 10 year olds. Hillis et al. (2002) found that when adults combined texture and disparity cues to slant, they lost access to the individual cue estimates such that they could not distinguish perceptual metamers (see Section 1.4.1). If children are not integrating the two cues to slant, then this mandatory fusion is unlikely and we would expect them to maintain access to the individual cue estimates of slant. Nardini et al. (2010) tested this in a second experiment, in which metameric stimuli were presented with equal and opposite slant offsets applied to the texture and disparity images. Observers made 'same or different' judgments for pairs of stimuli with varying combinations of texture and disparity slant. Adult observers could only discriminate between stimuli on the basis of a fused slant estimate: metamers could not be discriminated even though the changes in slant would have been distinguishable in the single cue case. Children (6 years old), however, did not gain the benefits of cue integration but as a result retained the ability to discriminate between metameric stimuli as well as they could discriminate between the slant estimates from individual cues.

No other study has examined the development of optimal cue integration within a single modality; however, there are some recent studies of the development of cross-modal cue integration. Gori, Del Viva, Sandini & Burr (2008) tested visual-haptic cue integration using two tasks: a size discrimination task ('which block is taller?') and an orientation discrimination task ('which bar is steeper?'). In each task one stimulus was presented, either visually, haptically or both modalities together; the other stimulus was then presented and the observer reported either which was taller or which bar was steeper. Adults and 10 year old children were able to discriminate between size better in the bimodal condition than in either of the unimodal conditions, in line with the predictions of statistically optimal cue integration. The evidence that 8 year olds are able to integrate optimally for size was somewhat equivocal: they weighted haptic and vision appropriately according to their relative reliabilities, however, the optimal integration model could not explain the variance in their responses. It is clear, however, that 5 and 6 year olds do not exhibit optimal integration strategies for size estimation; they showed a greater reliance on haptics than would be predicted by its reliability (as determined from unimodal trials). A similar picture was reported for the orientation discrimination task: adults showed optimal integration, there was no data reported for 10 year olds and the evidence for 8 year olds suggested that there was some integration but it was not yet optimal. Eight year olds, in common with younger children (5 and 6 years old) showed underweighting of haptics relative to its reliability, although 8 year olds to a lesser extent than younger children. However, the optimal integration model was a better fit to the response data for 8 year olds than a visual-only or haptic-only model. This data suggests that

optimal integration strategies for visual-haptic stimuli mature between the ages of 8 and 10 years old. For young children, haptics dominated bimodal size judgements whereas vision dominated bimodal orientation judgements. This occurred despite vision being the more reliable unimodal cue for both tasks. Gori et al. suggest that this may be due to the need to calibrate senses during childhood using whichever sense is the more suitable for the task, regardless of whether it has the better precision. In this case size is not something that can be estimated directly from the retinal image as there is a confound between size and distance in terms of size in the retinal image. Haptics provides a more direct estimate of size since the position of digits remains constant with distance. Conversely, orientation can be determined directly from the retinal image whereas for haptics it requires multiple inputs from either different fingers or a single digit over time.

If haptics is used to calibrate visual size judgements and vision is used to calibrate haptic orientation judgements this could explain why the two estimates are not integrated by young children: the information is better used for calibration than for noise reduction. To test this theory of cross-modal calibration, Gori, Sandini, Martinoli & Burr (2010) repeated the haptic size and orientation tasks described above with visually-impaired children aged 5-19 years old and age matched controls. They found that haptic size estimates in visually-impaired children were as good or better than the sighted control group. By contrast, haptic orientation estimates were significantly worse in visually-impaired children than in the control group, with a notable exception: one 10 year old child, rather than being congenitally blind, had normal vision until 32 months old and then became visually-impaired; this child had haptic orientation thresholds similar to age matched controls. These results support the hypothesis that vision is used to calibrate haptic orientation estimation. However, the child who lost vision at 32 months had normal haptic orientation thresholds, suggesting that the majority of haptic calibration can be completed in 2.5 years and may not need the extended period of time that Gori et al. (2008) indicate although it is hard to draw strong conclusions from a single observer. A further study by Gori, Tinelli, Sandini, Cioni & Burr (2012) tested children with motor disabilities (aged 5-18 years old) on the orientation and size discrimination tasks described above and compared their performance with age matched controls. Children with motor disabilities had visual orientation thresholds that were as good as the control group. However, they had significantly higher visual size discrimination thresholds than the control group. As in the previous study (Gori et al., 2010), they tested one 17 year old child who had acquired a movement disorder at 2 years old. This child's data showed normal visual discrimination thresholds for both tasks, again suggesting that the majority of calibration happens very early, perhaps before 2 years of age. Although the results of Gori et al. (2010) and Gori et al. (2012) support that idea that haptics is used to calibrate vision it is not clear that optimal cue combination precludes

continuing recalibration of perceptual cues since there is remarkable flexibility in adult observers to recalibrate cues (e.g., Adams et al., 2001, 2004, see also Section 1.6.1).

Ernst (2008) suggested an alternative possibility as to why children might not combine visual and haptic cue estimates: to integrate cues across modalities one must attribute the two cues to the same world property (a correspondence problem). In the tasks used by Gori et al. (2008) observers reached behind a screen to feel the haptic stimulus whereas the visual stimulus was in front of the screen. Although observers were informed that it was the same object on both sides, protruding through the screen, it may be that children did not associate the visual and haptic stimuli with one another. A similar task by Drewing & Jovanovic (2010) addressed this issue by showing a standard stimulus through lenses such that observers saw and/or felt one stimulus and compared its size to other (haptic only) stimuli. In the visual and visual-haptic conditions, observers could see their own hand feeling the object. In this scenario where the stimuli were co-located and size conflicts between the two modalities were small, a sub-optimal cue integration model predicted the data for adults and children (5-6 years old) better than either a switching model or an optimal cue integration model.

To avoid the problem of the two cue estimates being dissociated from one another, Nardini, Jones, Bedford & Braddick (2008) devised an object positioning task. Observers walked to and picked up 3 glowing objects within a darkened room where the only other visual landmarks were 3 illuminated shapes. Subsequently they attempted to replace the first object back in its original location. Three conditions were tested: vestibular only; visual only and bimodal (visual and vestibular). In the vestibular only condition, the three illuminated landmarks were switched off so that the observer had to rely solely on the vestibular estimate of object location. In the visual-only condition observers were spun on the spot before replacing the object such that they could rely only on the visual landmarks. In the bimodal condition the landmarks remained on and observers were not spun so that both estimates were available to determine where to place the object. Adult observers demonstrated optimal integration of the visual and vestibular estimates, replacing the object with greater accuracy (lower mean square error) and precision (smaller variance) than in either of the unimodal conditions: the reduction in variance was as predicted for optimal integration. By contrast, children (both 4-5 year olds and 7-8 year olds) showed no improvement in mean square error or bimodal variances, indeed these were not even predicted by using the most reliable single cue (visual landmarks) but were better predicted by an alternation or switching model in which responses were drawn variously from the two single cue estimate distributions. This supports the evidence from Gori et al. (2008) that cross-modal cue combination is not optimal until after 8 years of age. A potential drawback of the studies by Gori et al. (2008) and Nardini et al. (2008) is that they both require working memory: the former is a two-interval forced choice design in which the observer must remember how tall or steep the

previous stimulus was in order to compare the current stimulus; the latter requires observers not only to estimate position but also to remember that position.

Nardini, Begus & Mareschal (in press) tested visual and proprioceptive cue integration in a task not requiring memory. A location was marked on the top of a table and participants indicated its position with a finger using a touch pad on the underside of the table. In visual only trials, participants saw the marked location whereas in proprioceptive trials the participant held their finger at the marked location and then indicated position. In bimodal trials, observers both saw and touched the marked location whilst they indicated its position on the underside of the table. Adults and 7-9 year old children showed optimal integration: variances were reduced in the bimodal condition as predicted by a Bayesian model. However, 4-6 year olds and 10-12 year olds did not show the same improvement in variance as predicted by the optimal integration model (although some children did exhibit some benefit from using both cues). This is a surprising finding as it suggests that the maturation of cue combination may not follow a monotonic trajectory. The authors suggest this may be due to the onset of puberty, a time of rapid growth, requiring recalibration of proprioception. However, Nardini et al. (2010) found that 8 year olds benefitted from bimodal visual conditions compared with their least reliable cue in a visual-visual cue combination task whereas 6 and 10 year olds did not. It may be that puberty causes not only physical growth but also significant changes in perception.

The development of audio-visual cue integration has not yet been studied within the Bayesian framework. There is evidence to suggest that children have some capacity to use information from these two modalities concurrently. McGurk & MacDonald (1976) found that the auditory and visual components of speech were combined such that a video of lip movements saying 'ga' together with a soundtrack that played 'ba' was perceived as the syllable 'da'. Children aged 3-4 years old and 7-8 years old experienced this effect although not as strongly as adults. Subsequently this effect has also been found in infants (Rosenblum, Schmuckler & Johnson, 1997; Desjardins & Werker, 2004). Tremblay, Champoux, Voss, Bacon, Lepore & Théoret (2007) studied the McGurk effect across a number of age groups and found that 5-9 year olds integrated auditory and visual stimuli, but to a lesser extent than 10-14 and 15-19 year olds. Although the strength of the McGurk effect appears to be greater in older children and adults, it is not clear whether this is as a result of immature integration mechanisms or simply differences in the relative reliabilities of the cues at different ages.

Scheier, Lewkowicz & Shimojo (2003) used a version of the bounce/stream task (described in Section 1.4.2) to assess the capabilities of infants to integrate audio-visual events. Infants (aged 4, 6 or 8 months old) were habituated to a stimulus that showed two identical discs streaming past each other whilst a sound played either synchronised with the discs' coincidence or 1.3s offset from coincidence (either before or after coincidence). The test stimulus was then the offset or coincident stimulus respectively.

They found that 6 and 8 month old infants could discriminate between these stimuli (as shown by increased looking times) whereas 4 month old infants did not discriminate between them. The authors interpret this as evidence that 6 and 8 month old infants experience the illusory 'bounce' percept when the sound is concurrent with the discs' coincidence but experience streaming when the sound is offset. In a further experiment, Scheier et al. habituated infants to the sound coincident stimulus described above and tested looking times when presented with a new stimulus in which one frame was removed such that the discs no longer overlapped at coincidence but rather paused: they describe this stimulus as a 'physical bounce' as opposed to an 'illusory bounce'. Looking times indicated that infants did not perceive this as novel, which Scheier et al. take as further evidence of infants experiencing the illusory bounce percept. However, they did not test these stimuli without the coincident auditory stimulus so it is not clear whether infants could distinguish the visual components without the sound present. This does not, therefore, provide convincing evidence that the auditory stimulus changed the visual percept. Slater (2003) also suggested that the ability to distinguish between the coincident and offset auditory stimuli, in the first experiment described here, might reflect an ability to distinguish the relative timings of auditory and visual stimuli, rather than an altered visual percept due to changes in the relative timing of the auditory stimuli as claimed by Scheier et al. (2003).

Another approach to infant audio-visual integration was that of Neil, Chee-Ruiter, Scheier, Lewkowicz & Shimojo (2006) who compared reaction times of 1-10 month old infants to unimodal and bimodal stimuli. Stimuli were lights or sounds located either $\pm 25°$ or $\pm 45°$ from the midline. Orienting reaction times were measured for each stimulus combination (audio-only, visual-only and audio-visual). Bimodal reaction times were significantly faster than both unimodal conditions for 2-4 and 8-10 month old infants at 25° eccentricity and 0-2, 4-6 and 8-10 month old infants at 45° eccentricity. Bimodal reaction times for adults were significantly faster than both unimodal conditions at both eccentricities. Neil et al. (2006) compared the bimodal improvement to the improvement predicted by the 'race model' (Miller, 1982). In the race model the sensory inputs are not combined but rather whichever individual modality is faster on any given trial will trigger a response. If the reaction times for the unimodal stimuli are variable and the distributions overlap then the race model predicts that, on average, response times to bimodal stimuli will decrease compared with stimuli presented in either modality alone. This is stated mathematically by Miller (1982) as:

$$p(RT < t|S_a, S_v) \leq p(RT < t|S_a) + p(RT < t|S_v) \tag{1.12}$$

Neil et al. (2006) found that the decrease in reaction times in adults and 8-10 month old infants violated the race model inequality at both eccentricities, suggesting that

they combine auditory and visual stimuli. The only other group to violate the race model predictions were 0-2 month old infants at 45° eccentricity. The authors suggest that this latter finding was somewhat unreliable due to large individual differences between participants; they thus conclude that multimodal integration develops late in the first year of life. Although they show violations of the race model in adults and older infants, they do not test whether their findings are consistent with a Bayesian optimal integration model so it is hard to compare this finding with other multimodal integration studies.

There have been a couple of studies considering audio-visual integration in older children: Tremblay et al. (2007) tested children between 5 and 19 years old on the flash/beep illusion described in Section 1.4.2. They found no significant effect of age between three groups (5-9, 10-14 and 15-19 years old) in the number of fission or fusion illusions experienced, concluding that audio-visual cue combination develops before the age of 5 years old. However, Tremblay et al. did not test an adult control group, so could not tell whether the fission and fusion illusions were adult-like even in the oldest age group. The authors also did not test whether the rates of fission and fusion illusions experienced by their participants were more consistent with optimal Bayesian cue integration, or more simplistic models such as the 'switching' model used by Nardini et al. (2008). Innes-Brown, Barutchu, Shivdasani, Crewther, Grayden & Paolini (2011) carried out a similar experiment with an additional adult control group; they found that children aged 8-17 years old experienced more fission illusions than adults, but no more fusion illusions. Innes-Brown et al. concluded, in contrast with Tremblay et al. (2007), that audio-visual integration matures late in childhood. However, they did not measure the unimodal performance for auditory stimuli, so it is not possible to reject the alternative hypothesis that differences in audio-visual performance are due to optimal integration of unimodal cue estimates with differences in the relative uncertainty of those estimates in childhood. Although they collected reaction time data, the lack of auditory-only data also precluded testing of the race model predictions as in Neil et al. (2006).

The conflicting evidence from these studies suggests that audio-visual cue combination develops some time between 8 months and 17 years old; there is insufficient evidence to determine at what age audio-visual integration becomes statistically optimal. In Chapter 4, I test both the bounce/stream and flash/beep illusions in children (5-7 years old) and adults; for the latter experiment, I compare both Bayesian and switching models of cue integration to determine whether observers behave in a statistically optimal fashion.

1.6 Cue Recalibration and Learning

How sensory systems come to represent the relationships between world properties and sensory inputs is still an open question. Broadly speaking there are two possibilities: either they are innately specified or they are learnt during a person's lifetime. These possibilities are not necessarily mutually exclusive: it could be that some cues are hard-wired, learned over the evolutionary lifetime of the species, or have genetically encoded 'default' settings but are then adapted over a lifetime, whilst other cues could be calibrated against the hard-wired cues (Scholl, 2006). However they are acquired, cues must be continuously updated to reflect changes in the world statistics on which they rely. Section 1.6.1 reviews evidence for such cue recalibration; Section 1.6.2 explores to what extent adults can learn new cue relationships.

1.6.1 Cue Recalibration

In adulthood, as discussed in Section 1.3, when two cues are available the resultant estimates are combined. For sensory cues to remain useful they must also be continuously updated to reflect changes in environmental statistics or physical changes (e.g., growth, muscle fatigue or ocular changes), a process known as recalibration. It is possible that when the estimates from two cues are in conflict, rather than (or as well as) being combined, one (or both) cues could be recalibrated to better fit with the other estimate. A recalibration of this kind would be recognised by a change in the estimates from each cue presented individually. The ability to remain plastic and change the calibration of the cues would be useful so that in the case of injury or a change in the statistics of the world, perception does not remain biased. One example whereby the relationship between vision and cues from other modalities changes is when someone wears a new pair of glasses. It is the ability to recalibrate which means that after a short while of wearing the glasses and interacting with the world the wearer does not misestimate size or distance. Adams et al. (2001) tested recalibration of visual cues and found that changes in perceived slant after wearing a prism in front of one eye were not due to weight changes, i.e., down-weighting the (now) biased cue of disparity but rather the binocular disparity cues were recalibrated by adapting the mapping between disparity and perceived slant. Further evidence that cues can be recalibrated in adulthood is provided by Atkins, Jacobs & Knill (2003); they showed that haptic information could recalibrate binocular disparity cues to distance. The interpretation of stereo depth cues was measured in test trials; subsequently, in training trials, haptic feedback was discrepant from the visual cues. This led to the recalibration of stereo cues as measured in test trials after the training period. Similar haptic recalibration of visual information was shown by Adams et al. (2004): they showed that the light-from-above prior can be adapted by haptic feedback. They estimated individuals' light-from-above priors (see Section 1.2.3.1) by presenting a

series of ambiguous stimuli and assessing which looked concave and which looked convex. Participants were then given haptic feedback of shape in an environment in which the average lighting direction deviated from their original assumption by 30°. This had the effect that some of the shapes which had looked concave now felt convex and vice versa. After training with the haptic feedback their light priors were estimated again and were found to have changed in the direction of the trained lighting direction. This generalised to novel stimuli so could not be explained by participants learning the stimulus set[2]. In a similar study (Adams et al., 2010) the light-from-above priors were estimated and then trained with either haptic feedback, haptic and stereo-visual feedback or stereo-visual feedback. The haptic feedback had the same effect as described before, however the rate of learning was reduced when both stereo and haptic feedback were present: the learned shift in light prior was smaller than the haptic-only training condition. This might have been counter to expectations because the addition of stereo information increased both the amount and reliability of feedback. However, the addition of stereo feedback also removed the visual-haptic conflict: the shape of the surface was clear from the start of each trial and observers did not need to reinterpret the shape of the stimuli when they touched the surface. There are two possible explanations for the reduction in learning: (i) haptics may play a particularly 'special' role in the recalibration of visual cues and so it is visual-haptic conflict that drives recalibration; or (ii) there was no forced re-interpretation of the stimulus over time because the stereo information was always present: the initial interpretation was correct. The idea that touch is special and is used to teach or calibrate visual cues has a long history (Berkeley, 1709) and is intuitively appealing. To test between these two possibilities the third condition gave stereo feedback that was intermittent, in the same way that haptic information is intermittent since it is only available at the point one is touching. The shift in light prior in this condition was the same as when there was only haptic feedback, suggesting that the key factor in recalibration is the reinterpretation of the stimulus in a different way. Most of the time all visual cues occur simultaneously and so it would be unusual, but not impossible, for one visual cue to recalibrate another; in this sense, haptics could play a special role in visual recalibration.

1.6.2 Cue Learning

There is mixed evidence as to whether and how adults learn new perceptual cues. Michel & Jacobs (2007) drew a distinction between 'parameter learning' and 'structure learning'. Parameter learning involves modification or adaptation of the relationship of cues to properties in the world but only between relationships that already exist. In

[2]It is interesting to note that this ability to recalibrate the light-from-above prior is not shared across all species: chickens raised in an environment in which light always came from below retained a light-from-above prior, suggesting that this is 'hard-wired' (Hershberger, 1970).

this sense parameter learning is somewhat akin to recalibration (as described above) although the initial calibration may be that the two cues are uncorrelated but could plausibly have some relationship to each other. Michel & Jacobs trained observers with stimuli in which visual motion direction was correlated with an auditory cue (different frequency bands of filtered white noise). They predicted that observers would be able to learn the auditory cue to motion direction because movement, in a natural environment, often causes both visual and auditory stimuli so the two stimuli would be plausibly related, although uncorrelated previous to the experiment. Observers did learn this new cue relationship; Michel & Jacobs claim that this is evidence of parameter learning. Structure learning, by contrast, is defined as the learning of new cue relationships for which no ecologically valid causal mechanism exists. Michel & Jacobs tested several examples: they paired a disparity cue with motion direction; a luminance cue with motion direction; a disparity cue with lighting direction; and an auditory cue with lighting direction. In each case they argued that there would be no reason for such cues to be related and so learning of any of these relationships would constitute structure learning. They found that none of these relationships were learned by observers and concluded that whilst parameter learning is possible, structure learning is not possible in adulthood (they acknowledge that structure learning may be possible or necessary in infancy or early childhood). The problem with this hypothesis, as it stands, is that anything that is not learnt can be classed as structure learning whereas anything that is learned can be called parameter learning. The dichotomy does not produce testable hypotheses as this requires speculation as to which cue relationships are considered to be plausible and pre-existing. Moreover, the limited exposure to cues in the laboratory may mean that there was simply insufficient time to learn the cues which had previously been unrelated. These cues and the world property with which they are correlated have a long term historical correlation of zero or close to zero. Any perceptual system would be suboptimal to learn a new cue rapidly against such a weight of evidence that the cue and property are unrelated.

Haijang, Saunders, Stone & Backus (2006) also found differences in the ability of observers to learn new cues. They used an associative learning paradigm which pairs an unconditioned stimulus with a conditioned stimulus until the conditioned stimulus is able to elicit the same response as the unconditioned stimulus. To use this paradigm for visual cue recruitment requires that a percept be considered to be a response, initially an unconditioned response to the original cues and subsequently a conditioned response to the newly associated cue. A new cue is paired with existing cues that elicit a certain percept; if the new cue is learned then it will elicit the same percept even when the pre-existing cues are removed. Haijang et al. used a rotating Necker cube whose rotation direction was ambiguous, resulting in a bi-stable percept; they added depth cues to force the perceived direction of rotation. Direction of rotation was simultaneously associated with another new cue (either position, translation or sound). They found that the visual cues (position and translation) were both learned by

observers and acted as a conditioned cue to depth, as measured by direction of perceived rotation when the Necker cube was presented without any other depth cues. They refer to this process as 'cue recruitment'. Unlike the new visual cues, the auditory cue was not recruited as a cue to rotation direction. More recently, Jain, Fuller & Backus (2010) used the same rotating Necker cube stimuli but contrasted the ability to learn 'intrinsic' and 'extrinsic' cues, finding that whilst the former could be learned, the latter could not. They define intrinsic cues as cues that are generated by the same image elements as the object (e.g., position of the Necker cube) whereas extrinsic cues are not generated by the same image elements (e.g., surrounding annulus of dots or auditory stimuli). This difference in the ease of cue learning is similar to that described by Michel & Jacobs (2007) but offers a more testable and precise distinction between the cases.

Ernst (2007) used a different approach to study cue learning: he suggested that if a new relationship between cues was learnt then when both cues were present but in conflict with one another, they would be combined in a Bayesian way and so after training, observers would not be able to discriminate between stimuli that they could discriminate before training. He measured unimodal and bimodal discrimination thresholds for objects with varying luminance (visual cue) and stiffness (haptic cue). Observers were then trained in a situation in which stiffness and luminance were artificially correlated (either positively or negatively). If observers learned this relationship, Ernst suggested that bimodal discrimination performance should improve in test trials in which the luminance/stiffness relationship was congruent with the training condition, but deteriorate in test trials where luminance and stiffness were in conflict with the trained relationship. He found that observers could learn the new cue relationship and that it had the predicted effect on discrimination thresholds. He modelled the learned relationship as a 2-D coupling prior representing the strength of the certainty that the two cues are correlated and hence predictive of one another. A uniform coupling prior represents the case where the two cues are uncorrelated: any combination of the two world properties would be considered equally likely. At the other extreme, if the two world properties are perfectly correlated with zero variance then the prior also has zero variance and takes the form of a line. In between these two extremes Ernst (2007) models a prior for partially or weakly correlated cues as a Gaussian centred on the line of correlation, with the variance representing the level of uncertainty in the correlation. It is impossible to classify the learning demonstrated in this study as either parameter or structure learning: the coupling prior could have existed and been uniform before training (parameter learning) or the coupling prior could have been created through training (structure learning). The cues learned in this study do not appear intuitively to have any ecological validity but without extensive statistical measurements of real world properties it would be impossible to say for certain; this highlights the main weakness of the parameter vs. structure learning paradigm.

Di Luca, Ernst & Backus (2010) found further evidence that seemingly unrelated visual cues can be learned or recruited, even when the new cue is 'invisible' or rather, unnoticeable. They correlated vertical binocular disparity gradient (created by magnifying one eye's input vertically) with the direction of rotation of a cylinder. On training trials the new cue was paired with other depth cues that disambiguated the direction of rotation: horizontal disparity and occlusion. This method is comparable with that used by Haijang et al. (2006) but in this case observers were unaware of the new cue. They found that vertical disparity gradients were learned as a cue to rotation direction and the authors describe this as structure learning. They use a slightly different definition of structure learning to that described above, encompassing any shift from 'absence of dependency' (whether known by the perceptual system or not) to 'dependency' between cues; in other words, the pre-existence of a parameter specifying the correlation between two cues, such as a coupling prior (Ernst, 2006, 2007), would not render this parameter learning. Using this definition, it is not necessary to carry out statistical measurements of the correlations between world properties to identify structure learning; it is sufficient to assess whether observers know and utilise a dependency between two cues prior to training.

The cues recruited in the study by Haijang et al. (2006) could alternatively be considered to be contextual cues; when the rotating cube was in a particular location, or context, it elicited one percept and when in a different context it elicited another. This is consistent with findings from Atkins et al. (2003) who found that people were able to learn two calibrations of disparity cues, with each dependent on viewing distance. In their study, when objects were close, depth-from-stereo estimates were larger than when objects were further away. Contextual cues have also been found to be important in visuomotor calibration: Martin, Keating, Goodkin, Bastian & Thach (1996) trained people over a period of six weeks to learn two different gaze-throw calibrations. In training sessions participants threw a ball at a target whilst either wearing or not wearing prism glasses. When tested after training, participants had learnt to use the contextual cue of wearing the glasses to access an alternative calibration that was appropriate, even on the first trial of wearing the prisms and the first trial after wearing prisms. They also found that people were able to retain the two different gaze-throw calibrations for more than 27 months. In this case the glasses acted as a contextual cue to which calibration should be used.

Seydell, Knill & Trommershäuser (2010) described context specific cue learning in Bayesian terms as the ability to learn, store and selectively apply multiple priors for a single world property depending on the context. They used the cue-conflict paradigm (described in Section 1.3) to determine the relative weights applied to two cues to object slant: aspect ratio and binocular disparity. If the true aspect ratio of an object is known then its slant can be determined from the projected aspect ratio of the image on the retina; however, this cue requires prior knowledge or assumptions of the true

aspect ratio. Observers were trained in two contexts defined by object shape (either ellipses or diamonds). Each observer was trained such that in one of the contexts (either ellipses or diamonds) the aspect ratio was held constant at 1:1 (either circles or squares); in the other context, aspect ratio was randomly varied across training trials. In test trials, the aspect ratio and disparity cues to slant were conflicted by $\pm 5^{\circ}$ assuming that the true object aspect ratio was 1:1. From observers' responses the relative weights given to aspect ratio and disparity were calculated. Seydell et al. found that more weight was given to aspect ratio in the constant aspect ratio context than in the random aspect ratio context, suggesting that observers could learn multiple aspect ratio priors based on shape-defined context. However, when the experiment was repeated using object colour rather than shape as the contextual cue, observers did not weight the aspect ratio cue differently between the two contexts after training - even when explicitly told of the relationship between colour and aspect ratio. It is not clear why shape but not colour would be learned as a contextual cue; it is possible that shape and aspect ratio are more commonly related in the natural world (i.e., this is a more ecologically valid pairing), or that colour is not able to act as a contextual cue. In Chapter 5, I present a study investigating whether two different lighting direction priors can be learned and applied where context is specified by colour. This study uses visual-haptic stimuli to recalibrate the light-from-above prior as described in Adams et al. (2004, 2010) but extends this method to train two different calibrations, one for red stimuli and one for green stimuli.

1.7 Summary

The Bayesian approach has provided a strong theoretical framework, able to generate predictions as to how perceptual systems make use of sensory information. The formalisation of cues into likelihoods and priors has provided a common structure to consider cue combination, recalibration and learning both within a single modality and across multiple modalities. The Bayesian approach has extended our understanding of all these areas but there are many questions still outstanding. This thesis broadens our understanding in four key areas, applying the Bayesian framework where possible:

- Although shape and depth perception have been well researched using the Bayesian framework, less is known about material perception, especially cross-modal material perception. Chapter 2 expands our understanding of cross-modal material perception by considering the extent to which the visual system can make use of the broad range of sensory correlates available from other sensory systems. Specifically, it investigates the influence of haptic stimuli on a purely visual property: gloss.

- Gloss can be a cue not only to material perception but also to shape perception. Chapter 3 investigates whether the visual system generates a mutually consistent combination of shape, gloss and highlight disparity estimates; to test this, highlight disparity and the availability of shape cues are manipulated.

- Chapter 4 investigates the development of cue integration strategies in childhood. Since relatively little is known about material perception in adulthood, it is helpful to choose other topics that have previously been investigated in adulthood. To this end, Chapter 4 investigates audio-visual cue integration, testing, whether audio-visual integration happens and whether it is optimal (in a Bayesian sense) by 5-7 years old.

- Chapter 5 contributes to the literature on cue learning in adulthood by asking whether the visual system is able to learn and implement two context-specific priors for the same property. The light-from-above prior is one of only a few priors for which it is well established that they exist and that they can be recalibrated by a few hours of training (e.g., Adams et al., 2004, 2010). Whether multiple light priors can be learned for multiple contexts is, however, an open question.

Chapter 2

Does it feel shiny? Touch influences perceived gloss

Kerrigan, I. S., Adams, W.J. & Graf, E.W. (under revision for Current Biology) Does it feel shiny? Touch influences perceived gloss.

Experimental design, data collection, analysis and write-up were completed by Iona Kerrigan under the supervision of Wendy Adams and Erich Graf.

2.1 Abstract

Our perceptual system combines visual and haptic information to optimize estimates of 3D properties including slant (Ernst, Banks & Bülthoff, 2000) and size (Ernst & Banks, 2002). However, the integration of visual and haptic cues to material properties has been largely overlooked. In two experiments we show that gloss perception, a primarily visual property, can be modulated by the haptically accessible material properties of friction and compliance. Observers viewed a single object, rendered with or without a specular highlight. In visual-haptic trials, observers also 'felt' the virtual object, rendered to feel either soft and rubbery or hard and smooth. Observers judged whether the object was shiny or matte. We found that how an object feels affects its perceived gloss; objects that feel hard and smooth look glossier than those that feel soft and rubbery. Although friction and compliance are not reliable predictors of gloss, the visual system appears to know and use a probabilistic relationship between these variables to bias visual perception. This relationship is manifest in a flexible model of specular highlights: when touch suggests gloss, the visual system accepts highlights that deviate substantially from geometrically correct locations. In contrast, when the same visual object feels rubbery, spurious highlights are rejected: the object appears matte.

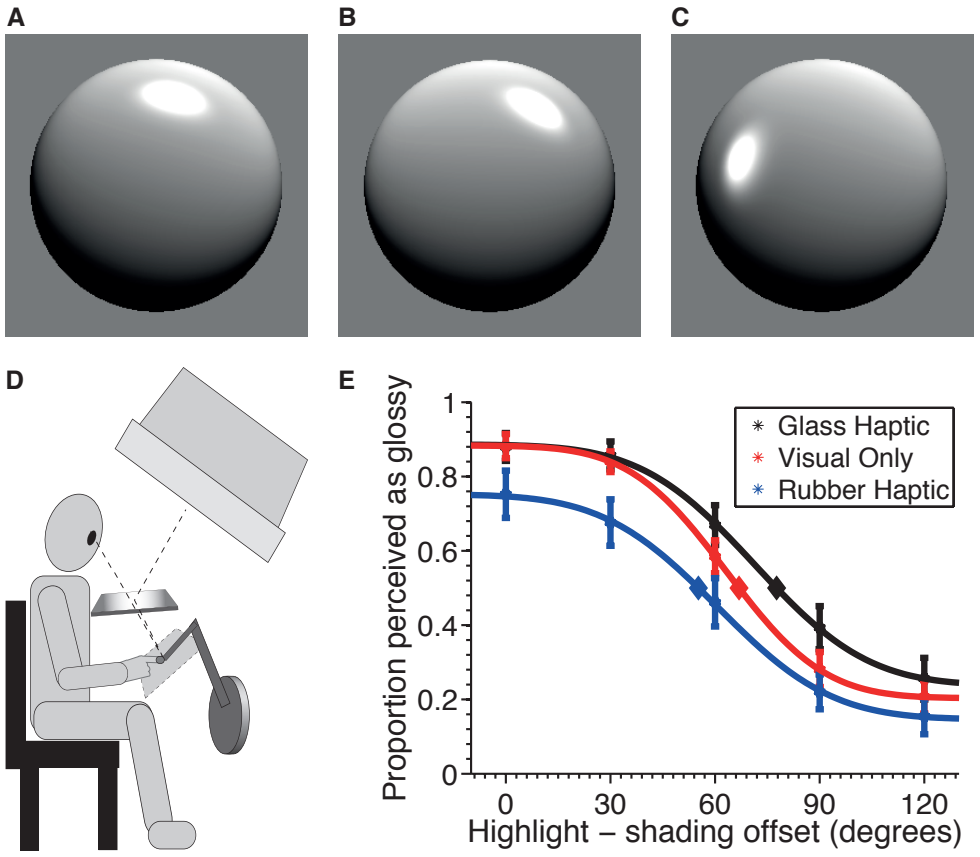

Figure 2.1: **Experiment 1: stimuli and results**
(A-C) Visual stimulus examples: tilt offsets shown are (A) 0°, (B) 30°, (C) -90°. (D) Visual-haptic set-up. (E) Data collapsed across shading gradient orientation and sign of shading-highlight offset; neither variable affected gloss judgements. Error bars give ±1SEM across observers. The 50% thresholds are indicated by large symbols.

2.2 Results & Discussion

Specular reflections occur when some or all light is reflected regularly from a surface, rather than scattered diffusely. These highlights give an impression of glossiness (see Figure 2.1A) and can thus be useful in identifying an object's material properties. Vision is not the only useful sense in such a task; the haptic (touch) properties of compliance, surface texture, thermal quality and friction (Lederman & Klatzky, 2009) can also contribute to material perception. How do these two modalities interact? We ask whether the perception of visual gloss is influenced by how an object feels. Despite the fact that gloss is conceived of as a visual feature, we show that the haptically accessible properties of compliance and friction influence how shiny an object looks. Specifically, haptic cues changed our observers' interpretation of visual highlights such that perceived gloss was systematically affected.

In one visual and two visual-haptic conditions, observers viewed a single smoothly

shaded disc, with or without a bright spot, and reported whether it appeared glossy or matte. In visual-haptic conditions, in addition to viewing the (virtual) object, observers simultaneously ran a finger over it, before responding (Figure 2.1D shows the visual-haptic set-up). In the 'glass' visual-haptic condition, the object had low compliance and low static and dynamic friction (like smooth glass). In the 'rubber' visual-haptic condition, the object had higher compliance and friction (like a squash ball).

Within each condition we manipulated the spatial alignment between the diffuse shading gradient and the specular highlight; see Figure 2.1. When the bright spot was spatially aligned with the bright area of the diffuse shading, it was interpreted as a highlight on a glossy object (Figure 2.1A). In both experiments, as the shading-highlight offset increased (Figures 2.1A - 2.1C) the object increasingly appeared matte (Figures 2.1E & 2.2); at larger offsets the bright patch was 'explained away' as a reflectance change or local spot of high illumination. This effect, formally evaluated in Experiment 2 (main effect of highlight offset: $F(4, 349.0) = 28.81$, $p < 0.001$), is consistent with previous demonstrations (Anderson & Kim, 2009; Beck & Prazdny, 1981). We used observers' tolerance to the shading-highlight offset to assess the effects of touch (Experiments 1 & 2) and highlight disparity (Experiment 2) on perception.

Haptic information significantly altered observers' visual percepts of material properties (Figure 2.1E). In the 'glass' condition of Experiment 1, objects were classified as glossy for significantly larger shading-highlight misalignments than in the visual-only condition ($p = 0.002$ from bootstrapping, after corrections for multiple comparisons). In contrast, visual-haptic 'rubber' objects appeared matte with significantly smaller shading-highlight offsets ($p < 0.001$). These results suggest that our perceptual system has an expectation regarding the glossiness of an object based on how it feels: hard and smooth objects are shinier than soft and rubbery objects.

In Experiment 2, we investigated how touch cues interact with specular highlight disparity. Unlike texture, highlights do not lie at the stereoscopically-defined surface depth. Instead, they lie behind convex surfaces, and in front of concave ones. Previous work suggests that observers implicitly know and use the geometry of specular highlights (Blake & Bülthoff, 1990, 1991), such that highlight disparity modulates perceived gloss. In Experiment 1, highlight depth (as defined by binocular disparity) was always geometrically correct: the highlight lay behind the convex surface. In Experiment 2 we introduced two additional highlight disparity depths: zero relative disparity (lying on the surface, like a paint spot) or reversed relative disparity (floating in front of the surface). As depicted in Figure 2.2B, we found that highlight disparity had a substantial and significant effect on gloss perception ($F(2, 647.2) = 69.8$, $p < 0.001$, from three-factor linear mixed model). Observers made the highest proportion of 'shiny' judgements when the highlight was behind the surface, at the

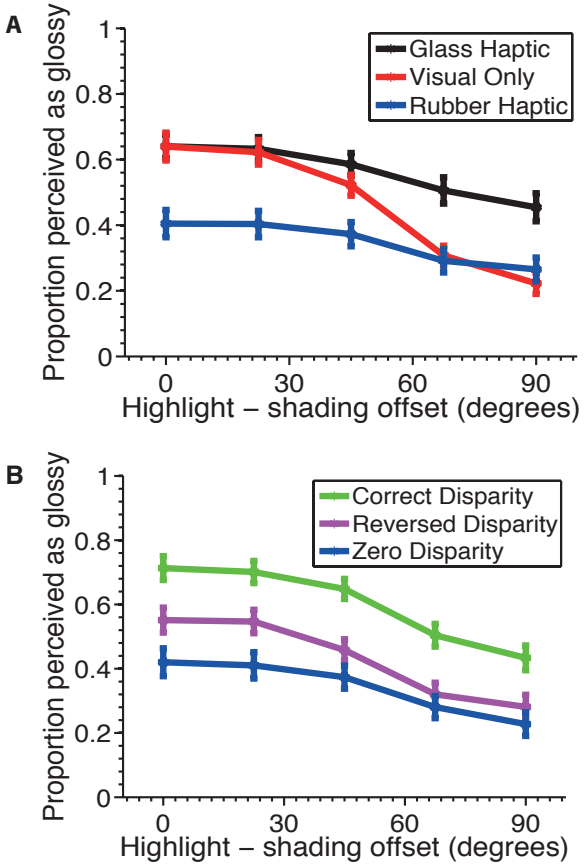

Figure 2.2: **Experiment 2: results**
(A) Stimulus categorisations as a function of highlight - shading offset, separated by visual-haptic condition and averaged across highlight disparities. (B) The same data, separated by highlight disparity and averaged across the visual and visual-haptic conditions. Error bars show ±1SEM across observers.

geometrically correct disparity-defined depth. When the highlight lay on the surface (zero relative disparity) the object was far less likely to be perceived as glossy. This makes sense: a bright patch at the surface depth is consistent with a local reflectance change, or a patch of higher illumination (e.g., from a spotlight). Interestingly, when the highlight was moved further toward the observer (in front of the surface), gloss percepts increased, although remained lower than for the correct highlight location (all pairwise comparisons significant, $p < 0.001$, after Bonferroni corrections). In other words, our observers were somewhat sensitive to highlight disparity. However, there were apparent perceptual failures - gloss percepts were promoted when the highlight was displaced from the surface depth in the incorrect direction (lying in front of the convex surface). Two factors may have contributed to this effect: first, observers may have an incomplete, or simplified model of highlight geometry (Kerrigan, Adams, Graf & Chang, 2011, see also Chapter 3). Secondly, uncertainty in object shape may have led to incorrect highlight interpretation (Blake & Bülthoff, 1990, 1991): shading cues were ambiguous, consistent with either a convex or concave surface under different

illumination directions, and the stereo cues that specified a convex surface were weak. Specular highlight disparity is itself a potential source of shape information (Blake & Bülthoff, 1990, 1991); a 'near' highlight is consistent with a concave surface, perhaps adding weight to a concave, glossy surface interpretation, despite reliable haptic information that the surface was convex.

We confirmed the main effect of haptics on gloss perception reported in Experiment 1. In Experiment 2, the haptic condition again modulated perceived gloss ($F(2, 616.3) = 43.28$, $p < 0.001$), with 'glass' and 'rubber' haptic information respectively increasing and decreasing 'shiny' responses, relative to the visual-only condition (all three pairwise comparisons significant at $p < 0.001$ after Bonferroni corrections). Interestingly, we also found an interaction between haptic condition and the shading-highlight offset ($F(8, 242.5) = 3.90$, $p < 0.001$). Observers were more dependent on shading-highlight offset as a gloss cue on visual-only trials; when haptic information about material properties was available, observers down-weighted other gloss cues.

In summary, we show that cross-modal influences are not constrained to the classic studies of location and shape perception previously reported. Instead, cross-modal interactions must also be considered within the domain of material perception. Specifically, we demonstrate that the haptically specified properties of friction and compliance have a substantial effect on visual gloss judgments. We also confirm previous findings that appropriate highlight disparity promotes gloss perception (Wendt et al., 2008, 2010), but suggest that observers are not as sensitive to relative disparity sign (highlights in front of versus behind a surface) as predicted by a full model of highlight geometry (Blake & Bülthoff, 1990, 1991).

Both experiments reported here show that the perceptual system incorporates an assumption that hard, smooth objects are more likely to be shiny than soft, rubbery objects, but does this have a sound ecological basis? The physical relationships between gloss, friction and compliance are complex: decreasing surface roughness at the micro level (e.g., by polishing) can both increase gloss and decrease friction. At this scale, smoothness modulates friction by altering the contact area between surfaces (Krim, 2002). Similarly, higher friction can occur between compliant surfaces as they deform to increase contact. However, friction and gloss can be unrelated in organic structures, for example, in the nanostructure that controls the low reflectance of moth eyes (Wilson & Hutley, 1982) or the glossiness of feathers (Maia, D'Alba & Shawkey, 2011). Additionally, the predominant determinant of friction for many solid surfaces may not be roughness, but adhesive forces between thin adsorbed films on solid surfaces (Krim, 2002).

One of the main behavioural advantages for a learnt relationship between friction and gloss may relate to identifying lubricant surface coatings. Whilst lubricants can be

powdery and matte, they are more often water or oil-based, and highly glossy. Observers appear to use gloss in assessing the slipperiness of a surface (Joh, Adolph & Campbell, 2006; Lesch, Chang & Chang, 2008) - there are clear advantages to identifying a wet, slippery floor. Thus, although friction and compliance may not be highly reliable predictors of gloss across natural objects, the visual system appears to have an implicit understanding of the probabilistic relationship between these variables, and use this to bias visual perception.

2.3 Experimental Procedures

2.3.1 Stimuli & Apparatus

Our visual-haptic set-up (Figure 2.1D) allowed us to simultaneously present visual and haptic information in the same perceived location on visual-haptic trials (the simulated object was 57cm from the observer's eyes). Observers wore stereo shutter goggles (CrystalEyes) and head position was maintained using a headrest.

The visual stimulus, rendered with the Phong lighting model (Phong, 1975) implemented in OpenGL, was consistent with a hemisphere squashed in depth by a factor of 2 and subtended a visual angle of 7°. Its smooth (Lambertian) shading gradient was generated by a single simulated distant lightsource, with 30° elevation (angle between the lighting vector and the screen). The tilt of the lighting vector (angle between the projected lighting vector and the screen's vertical axis) took one of five possible values per trial: 0, ±15°, or ±30°. The position of the highlight varied independently of the shading gradient; lighting tilt was changed before highlight rendering. In Experiment 1 the tilt offset between shading and specularity had one of nine possible values on each trial: 0°, ±30°, ±60°, ±90° or ±120°. The highlight depth (as defined by binocular disparity) was always geometrically correct: lying behind the convex surface. In Experiment 2, the nine possible values of the shading-specularity offset were 0°, ±22.5°, ±45°, ±67.5° or ±90°. We also introduced two additional highlight disparity depths: zero relative disparity (lying on the surface, like a paint blob) or reversed relative disparity (floating in front of the surface).

Haptic information was provided via a PHANToM force feedback device, with haptic stimuli generated using OpenHaptics (SensAble Technologies). The visual and haptic stimuli matched in size, shape and location, creating the perception of touching and viewing a single object. The object felt either hard and smooth (glass condition) or soft and rubbery (rubber condition). The former had lower compliance, static and dynamic friction than the latter.

2.3.2 Procedure

Observers gave informed written consent and the study was approved by the local ethics committee. Twenty-two observers (all naïve, 2 male, μ= 22.3 years) took part in Experiment 1 and 24 observers (22 naïve, 7 male, μ= 20.9 years) took part in Experiment 2. On visual-haptic trials the stimulus was only visible while observers touched the stimulus. Trials were presented in a pseudo-random order; each combination of stimulus values was presented once within each of four blocks.

2.3.3 Data Analysis

Data from Experiment 1 were analysed by fitting psychometric functions to the group data for each condition and comparing the resultant 50% thresholds (using version 2.5.6 of the psignifit toolbox for Matlab, which implements the maximum-likelihood method described by Wichmann & Hill (2001)). To identify significant differences between conditions, data were resampled 10,000 times, with replacement, within each condition and resultant thresholds were compared. Responses in Experiment 2 did not reach the 50/50 threshold in all conditions; data were analysed using a 3-way repeated measures linear mixed model in conjunction with pairwise comparisons.

Chapter 3

Highlights, disparity and perceived gloss with convex and concave surfaces

Kerrigan, I. S. & Adams, W. J. (under review at Journal of Vision) Highlights, disparity and perceived gloss with convex and concave surfaces.

Experimental design, data collection, analysis and write-up were completed by Iona Kerrigan under the supervision of Wendy Adams; Wendy Adams completed data-fitting and cross-validation analyses. Aaron Shuai Chang carried out data collection for the supplementary experiment and this data was submitted for his MSc dissertation at the University of Southampton; it has been analysed separately for this chapter.

3.1 Abstract

Glossy and matte objects can be differentiated using specular highlights: bright patches in the retinal image produced when light rays are reflected regularly from smooth surfaces. However, bright patches also occur on matte objects, due to local illumination or reflectance changes. Binocular vision provides information that could distinguish specular highlights from other luminance discontinuities; unlike surface markings, specular highlights lie not at the surface depth, but 'float' in front of concave surfaces and behind convex ones. I ask whether observers implicitly understand and exploit these peculiarities of specular geometry for gloss and shape perception. Participants judged the glossiness and shape of curved surfaces that included specular highlights at various depths.

Observers demonstrated substantial deviations from a full geometric model of specular reflection. Concave surfaces appeared glossy both when highlights lay in front of and

(incorrectly) behind the surface. Failings in the interpretation of monocular highlights were also apparent. Highlight disparity had no effect on shape perception. However, perceived gloss of convex surfaces did follow geometric constraints: only highlights at appropriate depths produced high gloss ratings. I suggest, in contrast with previous work, that the visual system invokes simple heuristics as gloss indicators to accommodate complex reflections and inter-reflections that occur particularly inside concavities.

3.2 Introduction

The objects and surfaces that we encounter daily are made of a wide variety of materials (e.g., stone, metal, plastic or fabric). Each of these materials reflect, refract and transmit light differently, enabling us to distinguish between them. In this chapter I consider the perception of surface gloss under binocular viewing. Whilst matte surfaces scatter reflected light in all directions, glossy surfaces reflect some (or all) light regularly, creating specular highlights. Both monocular and binocular cues can help observers to identify bright areas in the image that correspond to specular highlights. The 'orientation fields' of highlights provide one monocular cue: when a glossy object reflects its surroundings, these reflections are distorted according to the object's curvature. Unlike texture, highlights tend to 'cling' to areas of high curvature (Longuet-Higgins, 1960), such that they are aligned with long curvature axes (Fleming et al., 2004). In addition, it has been suggested that images of glossy objects have a characteristic skew in their luminance distribution that contributes directly to perceived gloss (Fleming, Dror & Adelson, 2003; Motoyoshi et al., 2007). However, skew alone is insufficient to promote percepts of gloss, spatial structure is also important (Fleming et al., 2003). The location and orientation of specular reflections must be consistent with the object's shape, as defined, for example, by diffuse shading patterns (Beck & Prazdny, 1981; Anderson & Kim, 2009; Kim et al., 2011; Marlow, Kim & Anderson, 2011).

Object or observer motion causes specular highlights to glide across the surface of a glossy object, rather than moving with it, like texture (e.g., Hartung & Kersten, 2002, 2003; Doerschner, Fleming, Yilmaz, Schrater, Hartung & Kersten, 2011). As binocular observers, we have access to analogous information even when the object is static, as we view the object from the separate vantage points of our two eyes. Highlight location depends on both the shape of the object and viewer position; a single light source is reflected from different surface points to reach the two eyes. Consequently, specular highlights do not lie at the stereoscopically defined depth of the reflecting surface. For simple curved surfaces, as shown in Figure 3.1, highlights lie behind convex surfaces, but in front of concave surfaces. (For specific, unusual viewing conditions, such as a light source lying between the surface and its focus point, these

rules can be broken, such that highlights appear behind concave surfaces.) With more complex surfaces and lighting, specular reflections generate disparity fields that may contain a wide variety of both horizontal and vertical disparities, whose corresponding depth field may contain discontinuities or be ill-defined (Muryy et al., 2012).

Previous work, exploiting simple curved surfaces, and clearly defined highlight disparity, suggests that observers use an accurate internal representation of highlight geometry when estimating gloss and shape (Blake & Bülthoff, 1990, 1991). More recent studies suggest that gloss perception is enhanced when highlight disparity (depth) is veridical (Wendt et al., 2008, 2010). Chapter 2 found that convex objects were interpreted as glossy more often when highlight disparity was veridical compared with when highlights had zero disparity (i.e., highlights lay on the surface). However, highlights with geometrically incorrect disparity sign (i.e., those lying in front of, rather than behind, the surface) were also interpreted as glossy more often than when highlights had zero disparity, although less often than when highlight disparity was veridical; these results would not be predicted if the visual system used an accurate model of highlight geometry. In addition, Blake & Bülthoff (1990, 1991) reported anomalous results for highlights on concave objects, not predicted by an accurate model of highlight geometry. They attributed these anomalies to rendering limitations producing conflicts between stereoscopic and monocular shape cues, such that shape was not reliably defined for concave objects. They concluded that humans:

> "employ a physical model of the interaction of light with curved surfaces...firmly based on ray optics and differential geometry".

I tested an alternative hypothesis: observers interpret highlights incorrectly on concave objects due to an incomplete model of highlight geometry. Observers reported both shape and gloss for surfaces with various highlight locations and shape cues of varying reliability (Figure 3.2). The results are clear: observers fail to make full use of highlight disparity when judging gloss on concave surfaces, even when shape is accurately perceived.

3.3 Methods

3.3.1 Subjects

Ten observers, including both authors, had normal or corrected-to-normal acuity, good stereovision (<40 seconds of arc, Stereo Fly test, Stereo Optical Company, Inc.) and no history of amblyopia or strabismus. Participants gave informed consent and the local ethics committee approved the study.

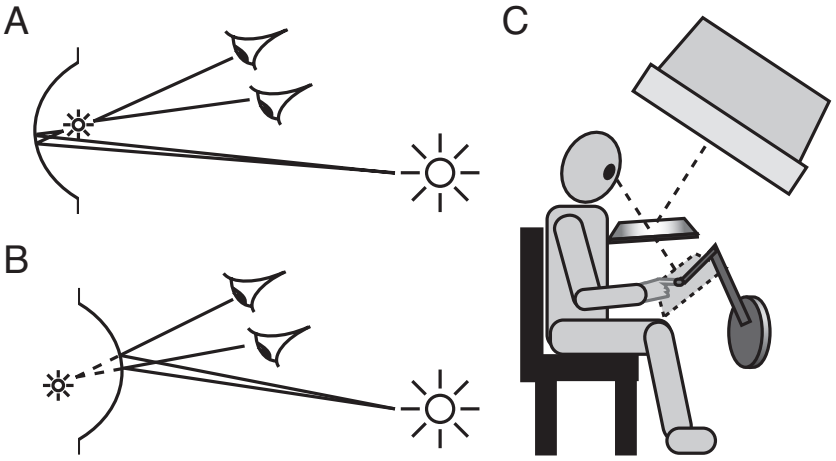

Figure 3.1: **Concave and convex highlight disparity and visual-haptic set-up.** (A & B) The image of a specular highlight appears in front of a concave surface (positive relative disparity) and behind a convex surface (negative relative disparity), adapted and reprinted by permission from Macmillan Publishers Ltd: Nature (Blake & Bülthoff, 1990), copyright 1990. (C) Visual haptic set-up: observers wore stereo shutter goggles (CrystalEyes) to enable stereoscopic presentation of stimuli; head position was maintained using a headrest. Haptic feedback was provided via a PHANToM force feedback device attached to the observer's right index finger.

3.3.2 Experimental set-up

Visual stimuli were generated using the Phong lighting model (Phong, 1975) implemented in OpenGL and displayed on a CRT monitor, viewed via a mirror as shown in Figure 3.1C. Haptic stimuli were generated using OpenHaptics and presented using a PHANToM force feedback device (SensAble Technologies). This set-up allows observers to view and touch simulated objects that are spatially aligned.

The study was conducted in a darkened room. Visual stimuli were consistent with one convex bump and one concave dimple on a plane (see Figure 3.2), illuminated by a single distant light source. The inclusion of both a convex and concave stimulus strengthens the shape percept for the concave object due to the single light source assumption (Kleffner & Ramachandran, 1992) counteracting the effect of the convexity prior (Langer & Bülthoff, 2001). The resultant shaded discs each subtended 6.6° at the viewing distance of 57cm, and were displaced ±4.2° horizontally from the screen's centre. Bumps and dimples were spherical sections, whose centres protruded or recessed 1.1, 1.6 or 2.4cm from the plane of the screen (Figure 3.3 shows cross-sections of these shallow, medium and tall objects). The lighting direction had a slant of 64° (the angle between the lighting vector and the screen normal) and a tilt of 135° (the angle between a horizontal axis in the screen's plane and the projection of the lighting

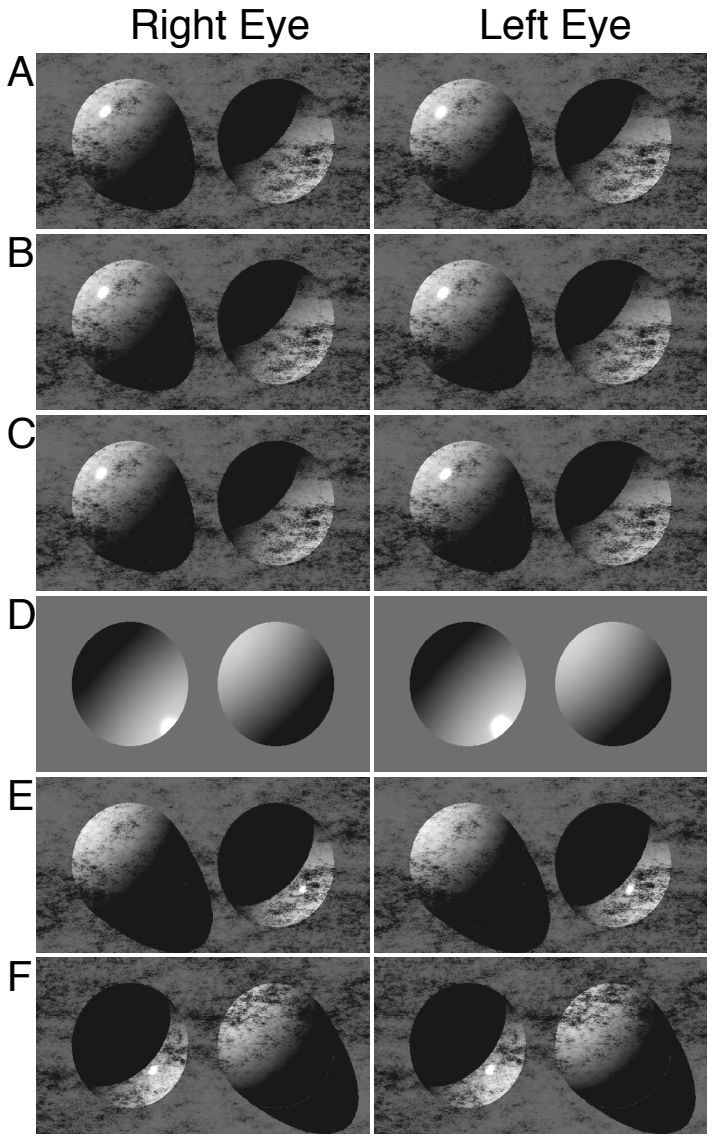

Figure 3.2: **Example stimuli stereo pairs for cross fusion.**
For demonstration purposes, example stimuli are presented here as stereo pairs for cross-fusion but were presented stereoscopically via shutter glasses during the experiment. (A-C) Visual stimuli for medium and high reliability conditions, 1.6cm height convex target object. Disparity of highlight is: correct; zero (on the surface); and reversed (i.e., in front of the surface), for A-C respectively. (D) Low reliability condition, 1.1cm height concave target, disparity of highlight is reversed (i.e., incorrect, as far behind the surface as a veridical highlight would have been in front). For concave objects at this shallowest depth (and not for any other objects) the highlight is close to or off the edge of the object for all disparities and thus has a different shape in the left and right eyes' images. (E) Medium and high reliability conditions, 2.4cm height concave target object. Disparity of highlight is -20 min arc (i.e., incorrect, at the furthest point behind the surface that was presented). (F) Medium and high reliability conditions, 2.4cm height concave target object. Disparity of highlight is correct.

vector). One of the objects had a specular highlight, whose position in depth relative to the surface (defined by its relative horizontal disparity) varied across trials, in the range −20 to +30 arc min. The highlight disparity values presented for each shape included the correct disparity, zero relative disparity, the opposite absolute disparity and the opposite relative disparity. The vertical disparity of the highlight was invariant, and geometrically correct. The horizontal disparity of the highlight was manipulated by changing the simulated eye positions used to generate the specular component of the stimulus. This allowed horizontal disparity to be manipulated independently of vertical disparity for the highlight, whilst keeping highlight shape and size consistent with the rendered surface. More complex illumination fields can enhance the perception of gloss (Fleming et al., 2003). However, despite a simple point light source, when the stimuli were viewed stereoscopically with the correct highlight disparity, authors and observers perceived very glossy surfaces that, in the medium and high reliability conditions, resembled polished marble. This combination of a simple surface shape and a single light source allowed us to systematically measure the effect of horizontal highlight disparity on gloss perception. To provide a 'matte' reference, a stimulus with no highlight was also presented, with a bump and dimple of ±0.8cm. At this depth, under the illumination and scene layout described, a glossy surface has no visible highlight.

On 'low reliability' trials, shape was defined primarily by shading, with a very weak binocular cue from the inter-ocular differences in the shading pattern (Figure 3.2D). In addition to shading, stimuli in the medium reliability trials were wrapped with a $1/f$ noise texture pattern and cast shadows were also rendered, providing more reliable depth information (see Figures 3.2A - 3.2C & 3.2E). High reliability trials were visually identical to the medium reliability trials, but included simultaneous and consistent haptic (touch) shape information. Observers are able to combine haptic and visual cues to reduce noise in perceptual estimates and disambiguate shape (e.g., Ernst & Banks, 2002; Adams et al., 2004; Helbig & Ernst, 2007b; Wijntjes, Volcic, Pont, Koenderink & Kappers, 2009). The scene was only visible whilst observers made haptic contact with it, ensuring that observers made their judgements based on the visual-haptic stimulus, rather than an initial, visual-only stimulus. Low coefficients of static and dynamic friction, and compliance were chosen such that the objects felt hard and smooth (implemented via OpenHaptics parameters: stiffness = 0.68; damping = 0.12; static friction = 0.28; and dynamic friction = 0.38).

3.3.3 Procedure

On each trial, an arrow indicated which of the two shaded discs should be judged. Observers adjusted a visual-haptic pointer on a sliding scale to indicate how shiny the object looked from 'not shiny' (0) to 'very shiny' (10). They utilised a wide range of

response values; surfaces without highlights were given low gloss values (gloss rating for concave and convex objects: $\mu_{concave} = 2.41$, SEM = 0.57, $\mu_{convex} = 2.10$, SEM = 0.55). Via another sliding pointer, observers adjusted a 2D contour until it matched the cross-section of the judged shape. Observers were given unlimited time to view the stimuli and adjust their settings before proceeding to the next trial. Each observer completed 2 blocks of 133 trials for the medium and high reliability conditions and 4 blocks for the low reliability condition. Each block comprised every combination of target object position (left/right), bump/dimple height (6 possible values) and highlight disparity (10 or 11 values, including monocular highlights and highlight absent condition). Trial order within blocks and block order across participants were randomised. The study took approximately two hours, including breaks.

3.4 Results

The manipulation of shape reliability was effective in modulating observers' curvature perception. With minimal shape cues (low shape reliability condition), the *sign* of surface curvature was accurately reported, but curvature magnitude was underestimated and responses were variable, particularly for concave surfaces (see Figure 3.3A). When reliable shape cues were available, observers' convex and concave surface perception was very accurate (medium and high shape reliability conditions, Figures 3.3B & 3.3C). Perceived shape was very similar in the medium and high reliability conditions, with no significant differences between the two except for the deepest objects (±2.4cm), where slightly more depth was reported in the high reliability condition ($F(1,7) = 5.70$, $p = 0.048$ ($\mu_{high} = 2.6$cm, $\mu_{medium} = 2.3$cm) and $F(1,7) = 27.97$, $p = 0.001$ ($\mu_{high} = -2.8$cm, $\mu_{medium} = -2.6$cm), respectively). In other words, the addition of haptic shape information had little effect on shape perception, suggesting that shape information provided by the visual cues in these conditions (shading, cast shadows, texture, disparity) was very reliable, allowing minimal effects of shape priors or residual cues to flatness (accommodation, vergence). Gloss ratings were also very similar across the medium and high reliability conditions, with significant differences only for the shallowest concave surface which was rated as slightly glossier in the medium reliability condition (LHS: $F(1,7) = 6.62$, $p = 0.037$, $\mu_{high} = 7.0$, $\mu_{medium} = 7.6$; RHS: $F(1,7) = 9.49$, $p = 0.018$, $\mu_{high} = 4.0$, $\mu_{medium} = 4.9$).

Despite excellent shape recovery in the medium and high reliability conditions, observers showed perceptual failures, relative to a full geometric model, when judging the gloss of concave surfaces (Figures 3.4A - 3.4C); larger symbols show the geometrically correct highlight depth for each surface shape - if an accurate model of highlight geometry were implemented, perceived gloss ratings would peak at these points. When highlights were located correctly in front of the surface, in agreement

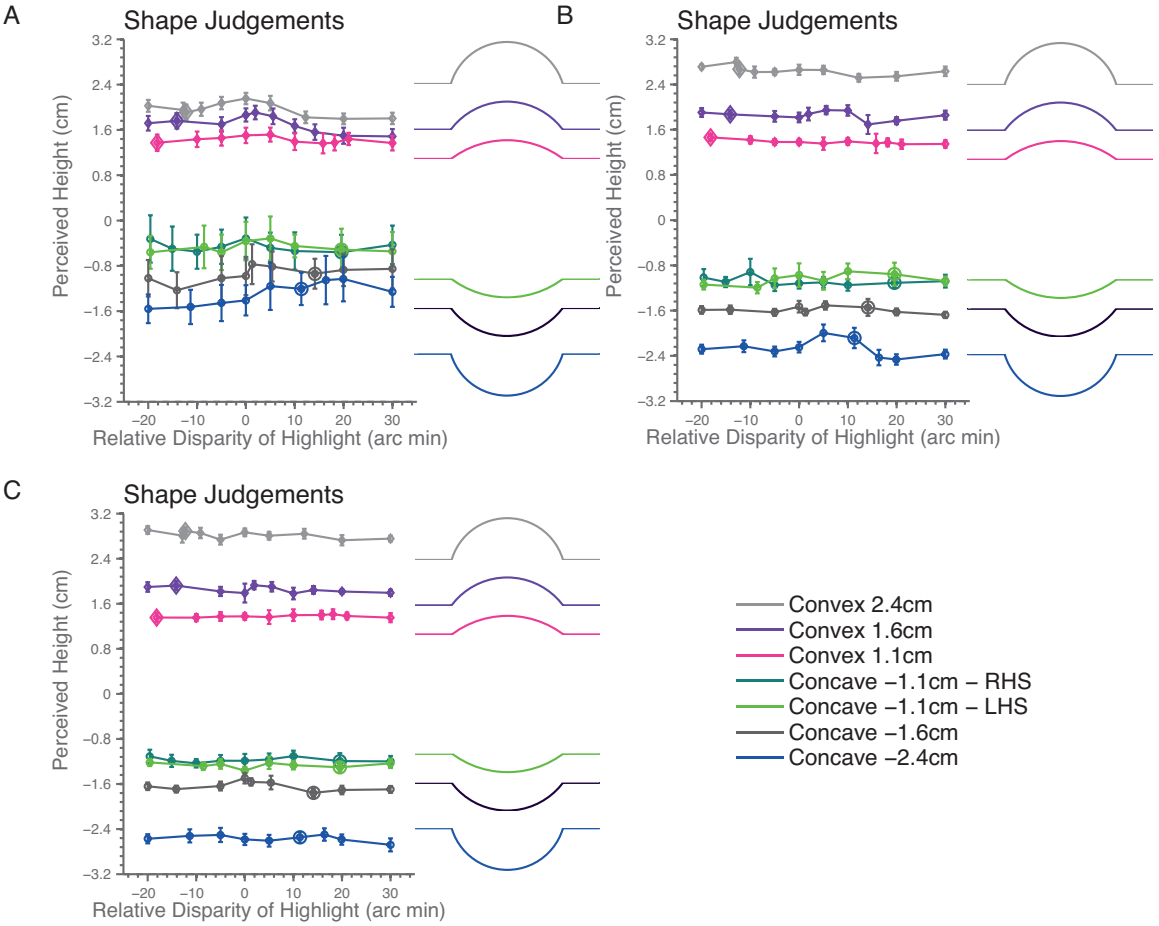

Figure 3.3: **Shape responses for low, medium and high reliability conditions.** (A-C) show shape responses for the low, medium and high reliability conditions respectively. Each shape is indicated by a different colour - see legend - error bars show ±1SEM. Adjacent contours show the true stimulus heights, each contour is aligned with the correct height response on the y-axis. Larger symbols show the geometrically correct highlight disparity.

with a geometrically accurate model, concave surfaces were perceived as quite glossy (average gloss rating for correct highlights across medium and high reliability conditions: 6.48). Also, as expected, gloss ratings were lowest when the highlight lay on the surface (zero relative disparity, average rating: 4.20); the straightforward interpretation of this case is a bright patch on a matte object caused by a light paint spot or local spotlight (Figure 3.2B). The dip in gloss ratings at zero disparity is clear across surface curvature sign and magnitude. Interestingly, however, as the highlight moved further away from its correct position, to sit behind the concave surface, that surface was again perceived as glossy, just as glossy as a surface with a correctly positioned highlight. In the 'behind' position, the highlight is far from correct, but observers appear to disregard highlight disparity *sign* for concave objects: both near and far bright spots are interpreted as highlights (average gloss rating with incorrect sign, but correct magnitude: 6.90). I tested sensitivity to highlight disparity sign by

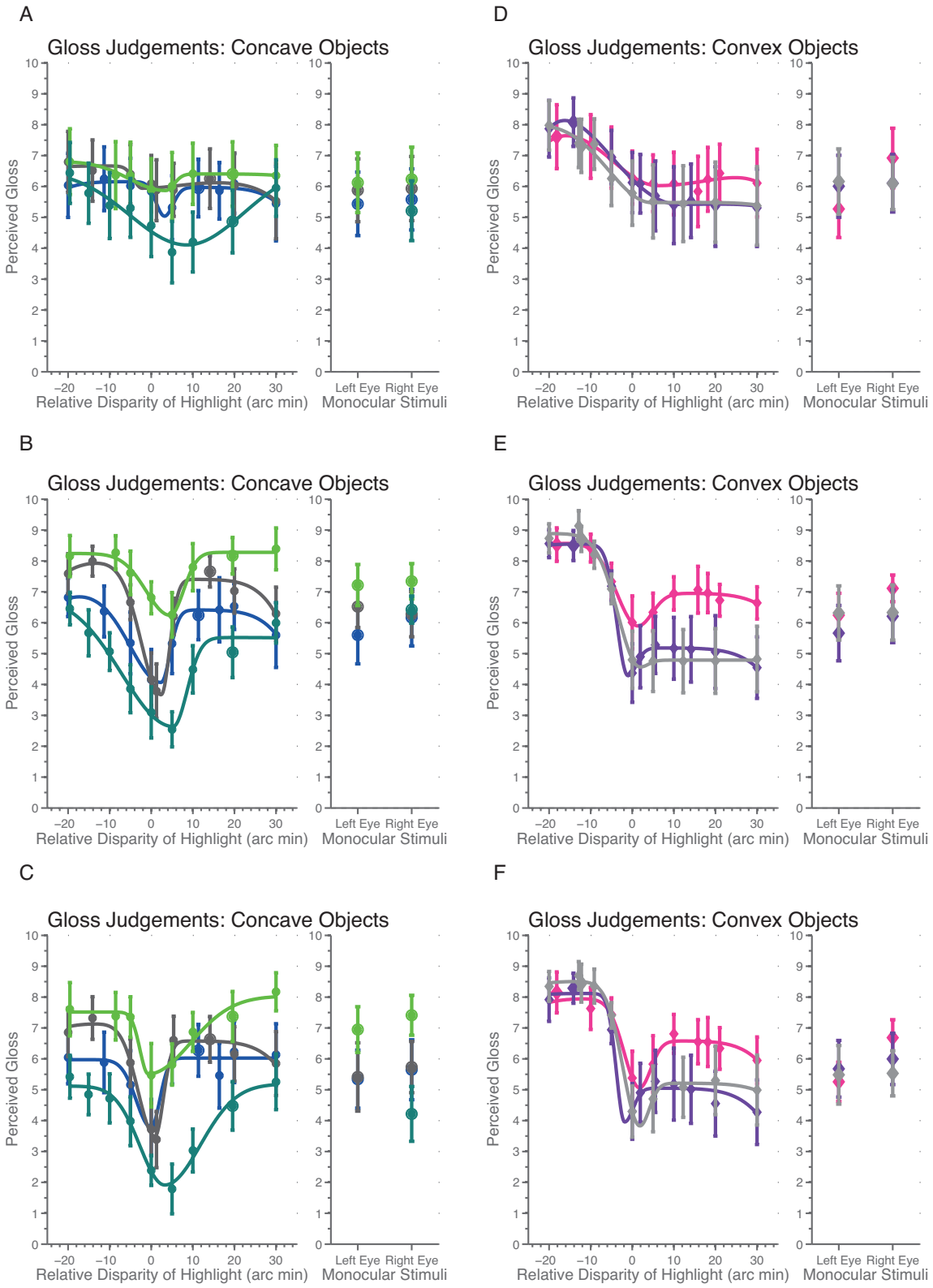

Figure 3.4: **Perceived gloss for low, medium and high reliability conditions.** See caption on next page.

Figure 3.4: (A-C) show perceived gloss for concave shapes in the three reliability conditions, respectively, whilst (D-F) show perceived gloss for convex shapes. Legend as in Figure 3.3, error bars show ±1SEM and larger symbols again show the correct highlight disparity. Gloss ratings for the shallowest concave surfaces have been plotted separately for objects on the left and right (light and dark green) because when this object was presented to the right of fixation the highlight was only partially visible to one eye due to the geometry of the scene, whereas when it was presented on the left, a large highlight was visible to both eyes.

comparing asymmetric and symmetric curve fits to the gloss rating data; data following a symmetric pattern would suggest that observers are insensitive to highlight disparity sign. The asymmetric fit was defined by an inverted Gaussian with five free parameters (the mean, corresponding to the low point of the function, and separate spread and scaling parameters on each side of the mean). For the symmetric curve fit, a single pair of spread and scaling parameters defined the curve on both sides of the mean. Responses to the medium and high reliability stimuli for concave shapes were equally well fit by the symmetric model, with the asymmetric model producing only a modest improvement (reduction in sum of squared residuals: $\mu_{medium} = 26\%$, $\mu_{high} = 21\%$, averaged across object depths). A cross-validation technique revealed that the asymmetric model was not significantly better than the symmetric model ($p > 0.05$ for all concave objects, medium and high reliability conditions, from t-tests on residuals calculated by excluding each data point in turn). This suggests that observers disregard the sign of highlight disparity, despite being sensitive to its magnitude. In fact, some observers appear to lack even a basic model of highlight geometry, despite good stereoscopic vision: two observers (excluded from all analyses) were completely insensitive to highlight disparity, perceiving any surface with a highlight as very glossy, irrespective of highlight depth.

In contrast with concave objects, a more sophisticated model of highlight geometry predicts observers' gloss perception of convex objects. Across all observers, convex objects appeared highly glossy when the highlight disparity was correct or close to correct (see Figures 3.4D - 3.4F, average gloss rating for correct highlight across medium and high reliability shape conditions: 8.44). The lowest gloss ratings were reported when the highlight had zero relative disparity (average gloss rating across medium and high reliability conditions: 4.86), and ratings were significantly higher for correct than incorrectly signed highlight disparity (mean rating for reversed highlight disparity: 5.58). I again tested sensitivity to highlight disparity sign using symmetric or asymmetric data fits. In the medium and high reliability conditions, where gloss was strongly modulated by highlight disparity, responses to convex objects were poorly fit by the symmetric model, with considerably better asymmetric fits (reduction in sum of squared residuals: $\mu_{medium} = 53\%, \mu_{high} = 50\%$). Cross validation confirmed that the asymmetric model provided a better fit for convex objects (all $p < 0.05$, except 1.6cm and 2.4cm objects in medium condition: $p = 0.06$ and $p = 0.08$). However, note the

larger error bars for the right hand side of plots 3.4E and 3.4F - some observers were more sensitive than others to disparity sign; a small sub-group of observers had a tendency to perceive convex surfaces with a near (incorrect) highlight as glossy. This was rather surprising given the alternative, viable stimulus interpretation: in the case of convex objects, a highlight with positive (incorrect) disparity can be perceived as a separate white object, like a small cloud, floating in front of a matte surface (see Figure 3.2C). This description fits both informal observations of the convex stimuli and the majority of the observers' data. Further analyses confirm that observers are more sensitive to incorrectly positioned highlights for convex, than for concave objects: averaged across disparity magnitude, convex objects with highlights (incorrectly) in front of the surface are perceived as significantly less glossy than concave objects with incorrect far highlights ($t_7 = 2.5$, $p < 0.05$; $\mu_{convex} = 4.9$, $\mu_{concave} = 6.5$, averaged across medium and high reliability and all convex or concave shapes, excluding the shallowest surfaces, where gloss is complicated by partially occluded highlights - see below). Furthermore, gloss ratings depend on disparity sign for convex, but not concave objects (main effect of disparity sign for convex objects: $F(1,7) = 8.1$, $p = 0.025$; $\mu_{correct} = 8.0$, $\mu_{incorrect} = 5.4$; but not concave objects: $F(1,7) = 4.2$, $p > 0.05$; $\mu_{correct} = 5.9$, $\mu_{incorrect} = 6.4$).

Across both convex and concave surfaces, highlight disparity had a smaller modulatory effect on gloss perception in the 'low' reliability condition, where the depth of the surface, and thus the relative depth of the highlight, was not reliably defined (Figures 3.4A & 3.4D).

For all highlight depths, gloss perception of concave surfaces was modulated by the magnitude of surface curvature: surprisingly, highly curved surfaces produced lower gloss ratings. The exception is the shallowest concave surface on the right of fixation (Figures 3.4A - 3.4C lowest line) - in this configuration only a sliver of a highlight was visible on the edge of the surface, and observers rated this surface as least glossy. I implemented the Phong model (Phong, 1975) to produce highlights whose spread is modulated by surface curvature (compare Figures 3.2A and 3.2D). The shape-gloss relationship in the data suggests that observers perceive surfaces with larger highlights as glossier, consistent with previous and recent evidence (Beck & Prazdny, 1981; Anderson, Marlow & Kim, 2012). This is true for gloss judgments of both concave and convex surfaces (main effect of highlight size, as indexed by object shape: concave surfaces $F(3,21) = 7.5, p = 0.018$, G-G correction, $\epsilon = 0.44$; convex surfaces $F(2,14) = 7.5, p = 0.051$, G-G correction, $\epsilon = 0.56$). The shallowest concave shape (1.1cm), presented on the right of fixation (small highlight) was perceived as less glossy than the same shape presented on the left (large highlight, $p = 0.002$); it was also less glossy than the 1.6cm concave surface ($p = 0.011$, Bonferroni corrected comparisons). The shallowest convex surface (large highlight) appeared glossier than deeper surfaces (1.1cm vs. 1.6cm, $p = 0.051$). This shape effect for convex surfaces was more

pronounced with incorrectly positioned highlights (interaction between highlight sign and shape magnitude $F(2, 14) = 7.3, p = 0.029$, G-G correction, $\epsilon = 0.52$); it appears that the 'larger highlights = more glossy' heuristic appears to come into effect when there is uncertainty in the interpretation of a potential highlight (concave objects, convex objects with misplaced highlights). There was also an interaction between highlight disparity sign and shape magnitude for concave objects ($F(3, 21) = 3.4$, $p = 0.036$); the modulation of gloss perception by highlight disparity is slightly different for the shallowest concave object presented on the right of fixation (bottom line, Figure 3.4C).

Why might the visual system use highlight size as a gloss cue? Smoother surfaces actually create smaller, more focused highlights, given a single distant light source. However, in the more complex light fields that typically illuminate natural scenes, very glossy or mirrored objects reflect the surrounding scene more sharply than matte objects (Pellacini, Ferwerda & Greenberg, 2000) and so may produce multiple large, bright areas.

Further limitations to the observers' model of highlight geometry were revealed when they viewed monocular highlights (within otherwise binocular scenes). Monocular highlights occur in normal viewing of glossy surfaces such as the stimuli used in this study, either when the surface does not extend far enough to 'capture' one eye's highlight, or when the light's path is occluded by the surface (the potential reflection point lies within a shadow). For a given scene it is geometrically plausible to have a monocular highlight visible to one eye, but not to the other eye. However, observers do not display this sensitivity, making equal, moderately glossy judgements for a monocular highlight presented to either the correct or incorrect eye (right hand component of each plot in Figure 3.4).

3.5 Discussion

Observers' interpretation of specular highlights reveals a limited geometric model. Although the glossiness of convex surfaces is perceived broadly in line with the expectations of an accurate geometric model, this is not the case across all stimuli: bigger highlights signal more gloss; monocular highlights viewed with either eye suggest moderate gloss. Surprisingly, for concave surfaces, highlights in entirely the wrong depth location - behind the surface - signal high gloss.

When the visual system is presented with a highlight at the wrong depth location (e.g., behind a concave surface rather than in front of it) it must find some 'explanation' of the retinal input. One interpretation, given a misplaced highlight, would be to amend the estimate of surface shape to accommodate it, e.g., a concave surface with a far (incorrect) highlight could be perceived as glossy and convex. Although

convex/concave reversals driven by highlight disparity have been reported elsewhere (Blake & Bülthoff, 1990, 1991), highlight disparity had little effect on perceived shape in the current study. It may be that a prior for overhead lighting (e.g., Kleffner & Ramachandran, 1992; Adams et al., 2004) and weak disparity cues remained strong enough to veto a reversed curvature interpretation even in the 'low reliability' condition. In a supplementary experiment I explored this possibility further by making object curvature more ambiguous with simulated lighting from the left or right. However, perceived curvature sign continued to be unaffected by highlight disparity - see Appendix 3.A for details. Rather than changing their shape estimate, observers perceived concave surfaces to be glossy, even when the binocular highlights were inconsistent with that interpretation. This failure to reject an interpretation of gloss was not due to uncertainty in shape as suggested by Blake & Bülthoff (1990, 1991): even when shape information was reliable and shape estimates were accurate, observers perceived concave surfaces with incorrect highlights as glossy.

A potential interpretation of an errant bright spot positioned behind a concave surface is of a light source viewed through a transparent surface. Is it possible that observers perceived not gloss, but transparency? Gloss and transparency are somewhat related: the separation of image components at different depths, corresponding to reflections and surface texture, is similar to that necessary to perceive transparency (Anderson, 2011) and gloss has even been conceptualised as a form of transparency (Mulligan, 1993). Furthermore, transparent objects are often glossy (Fleming & Bülthoff, 2005); the surface structure required to transmit light regularly will also result in mirror-like reflection of any reflected light. However, it is unlikely that observers perceived transparency from the experimental stimuli: casual inspection of Figure 3.2E reveals a percept of gloss rather than transparency; other transparency cues are absent - the surface in front of the light source is not visible with reduced contrast, and nothing else can be seen through the surface. Instead, it seems that the visual system determines that the bright spot in the image is a specular highlight on a glossy, concave object. Why might the visual system accept this interpretation? Firstly, if a light source is very close to a concave surface (between the surface and its focal point) a virtual, far image is produced. For the stimuli used here, such a light source would need to be less than 1.7cm or 2.7cm from the deepest or shallowest concavities respectively, and closer still to produce highlights at the actual disparity-defined depths presented to observers. Such light sources (unlike the simulated distant one) would be clearly visible in the image. Alternatively, light sources at low elevation (close to the image plane), could in theory produce inter-reflections leading to spurious binocular matches. Whilst these interpretations are inconsistent with the stimulus images, the visual system apparently knows that the relationship between highlight disparity and gloss is more complex for concave objects, and thus regards a highlight on a glossy surface as the most plausible explanation for the errant bright region. Observers thus perceived

highlights either in front of, or behind concave surfaces as indicative of gloss. Only highlights lying at the surface depth result in a perception of a matte surface.

Observers did implement the constraints of highlight geometry when assessing the glossiness of convex surfaces. This could be because highlight geometry is simpler for the convex than for the concave case: for convex surfaces, highlight images will always be virtual, located behind the surface, and not subject to surface inter-reflections. In contrast, given particular viewing and lighting conditions, highlights can appear either in front of or behind a concave surface, or can be partially or completely occluded from view by the surface itself. It appears that observers invoke a simplified model of highlight geometry that is robust to the complexities of reflections and inter-reflections inside concave objects, perceiving any potential highlight that lies off the surface (i.e., with non-zero relative disparity) as evidence of gloss. Additionally, due to the statistics of objects in our environment, observers may have had more experience with convex objects than concave, and may thus have more reliable representations of their highlight geometry.

Despite a simplified model of gloss, we rarely make errors in our estimates of object material. No doubt other cues to surface shape and gloss are exploited, such as the alignment of specularities and diffuse shading (Anderson & Kim, 2009; Beck & Prazdny, 1981), the relationship between highlight colour and surface colour (Nishida, Motoyoshi, Nakano, Li, Sharan & Adelson, 2008), flow fields of local image orientations in richer light fields (Fleming et al., 2004) and highlight motion (Hartung & Kersten, 2002, 2003; Wendt et al., 2010; Sakano & Ando, 2008; Doerschner et al., 2011). However, it remains surprising that the binocular disparity of specular highlights, which could be such a valuable cue to surface gloss and shape, constrains perception in an incomplete manner.

3.A Appendix

In a supplementary experiment, we investigated whether highlight disparity would have an effect on perceived curvature sign when shape was more ambiguous. To this end, simulated illumination was from the left or right (lighting tilt = 0° or 180°), rather than from above-left (consistent with a light-from-above prior). In addition, shading and texture were consistent with an object height of ±1.1cm, but the disparity of this shading and texture was consistent with a flat surface (as in Blake & Bülthoff (1990, 1991)). Observers estimated surface shape (via a sliding pointer, as in our main experiment) on trials with a 'concave highlight' (positive highlight disparity, consistent with a concave interpretation), 'convex highlight' (negative highlight disparity, consistent with a convex object) or no highlight.

In contrast to previous findings (Blake & Bülthoff, 1990, 1991), highlight disparity had only a small effect on the *sign* of curvature; all objects were perceived as convex on the majority of trials (proportion perceived as convex: 70.7%, 86.9% and 73.2% for 'concave', 'convex' and no highlight conditions, respectively). An ANOVA revealed a significant effect of highlight condition ($F(2, 26) = 4.73$, $p = 0.032$, G-G correction, $\epsilon = 0.70$), but this was driven by a significant difference between the 'no highlight' and 'convex highlight' conditions (Bonferroni pairwise comparisons, $p = 0.028$, other comparisons n.s.). Thus, we failed to replicate Blake & Bülthoff's finding that highlight disparity sign *determined* perceived curvature. Disparity sign did affect quantitative depth: the magnitude of depth estimates varied significantly according to the highlight condition (concave highlight: 1.12cm, convex highlight: 1.35cm, no highlight: 1.06cm, $F(2, 26) = 12.21$, $p = 0.001$), with significant differences between positive ('concave') and negative ('convex') highlight disparity ($p = 0.006$) and between the 'convex' and 'no highlight' conditions ($p = 0.007$, Bonferroni pairwise comparisons).

In summary, therefore, our supplementary experiment found that highlight disparity has little effect on perceived curvature sign, and a small but significant effect on curvature magnitude. There was an effect of highlight presence, with the 'convex' highlight producing more convex responses than no highlight. This finding is in agreement with a bias reported elsewhere (Bouzit, Adams & Graf, 2007); the presence of a highlight acts as a cue to convexity because highlights are more likely to be occluded on concave objects. Why did we fail to find the large effects of highlight disparity on perceived curvature sign, as reported previously? In contrast with previous studies (Blake & Bülthoff, 1990, 1991), our observers estimated quantitative depth, rather than making a 2AFC on convexity sign. This methodological difference, alongside stimulus differences may account for the somewhat disparate findings (our lack of any significant effect on curvature sign). Their stimuli showed convex and concave regions within the same object, so a convexity prior (e.g., Langer & Bülthoff, 2001) would have little effect on perceived curvature sign. In contrast, our stimuli were separated spatially which could have increased the influence of the convexity prior on the attended object (van Doorn, Koenderink & Wagemans, 2011). Together, these results suggest that whilst highlight disparity sign can be used as a cue to object shape, it is a very weak cue and will normally be dominated by other cues or biases, such as the convexity prior.

Chapter 4

Development of audio-visual integration

Experimental design, data collection, analysis and write-up were completed by Iona Kerrigan under the supervision of Wendy Adams. Fiona Berry, Nesta Caiger, Emma Ryan and Katie Hobbs helped with data-collection for these experiments and the data from both experiments were submitted for their BSc dissertations at the University of Southampton; they have been analysed separately for this chapter.

4.1 Introduction

To maintain a stable, unified percept of the world, it is helpful to combine information from different sensory cues. Shape and depth information is combined both within a single sensory modality (e.g., Hillis et al., 2004; Knill & Saunders, 2003) and across multiple modalities (e.g., Ernst et al., 2000). By combining multiple information sources it is possible to reduce uncertainty in estimates of world properties - for example combination of visual and haptic estimates of slant gives rise to a more precise estimate (Ernst & Banks, 2002). Chapter 2 and Chapter 3 considered how a similar approach can be applied to material perception. This chapter investigates whether optimal cross-modal integration is developed in 5-7 year old children. Since research into optimal integration for material perception is still in its infancy, an audio-visual task was chosen to investigate the development of cross-modal integration; this has the advantages that adult audio-visual integration has previously been characterised as optimal (e.g., Alais & Burr, 2004; Shams et al., 2005b) and that the equipment required to present stimuli could be transported easily to the primary school where the children were tested. Audio-visual cue integration has been shown to give rise to a variety of interesting illusions which are described below.

The ventriloquist effect occurs when a sound appears to originate from a different location to its actual source, due to a synchronous, spatially separated visual stimulus. Early studies attributed this effect to visual 'capture' of the auditory stimulus such that vision dominates the percept (e.g., Pick et al., 1969). The idea that vision is always 'superior' to audition was subsequently disproved in demonstrations of apparent 'capture' of vision by other senses under certain conditions (e.g., auditory capture of vision, Morein-Zamir et al., 2003). The assumption that vision will always capture audition arose because vision is usually superior to audition for spatial tasks, however, for temporal tasks audition is usually better and so it appears that there is auditory capture of vision (Morein-Zamir et al., 2003). Alais & Burr (2004) showed that the theories of visual or auditory capture were somewhat limited as they can only describe the extreme cases where vision or audition is much more reliable than the other. They modulated the reliability of visual stimuli in a localisation task and were able to recreate both auditory and visual 'capture' of the other sense; they also demonstrated that when the reliabilities of auditory and visual stimuli were similar, neither sense dominated and responses were an average of the two estimates. They explained their findings as optimal Bayesian integration of auditory and visual cues, each weighted in proportion to its relative reliability. The different theories of cue combination are discussed in more detail in Section 1.3.

Another example of visual percepts being altered by the addition of an auditory stimulus is demonstrated in the 'bounce/stream' effect (Sekuler et al., 1997). In this illusion, two identical discs move towards one another and continue past each other. Most adult observers perceive this as two objects moving past each other ('streaming'). However, the addition of a sound at the objects' coincidence results in a changed percept in which the objects appear to 'bounce' off one another. Whilst the bounce/stream effect has been widely studied in adults, only one study has examined it in childhood (Scheier et al., 2003). Scheier et al. (2003) used looking-time measures to determine whether infants experience this illusion; they found that infants from six months of age could discriminate between trials in which the sounds were presented either at the point of the discs' coincidence or offset in time by 1.3 seconds (the visual stimuli remained identical in both conditions). They attributed this to a percept of bouncing when the sound was coincident, and a percept of streaming when the sound was offset. However, looking time measures can be very difficult to interpret and Slater (2003) proposed an alternative explanation in which infants might experience the same visual percept in both conditions (sound coincident and sound offset) but respond to differences in audio timing relative to that percept.

The visual scene in the bounce/stream effect is ambiguous in the sense that both interpretations (bouncing or streaming) are consistent with the visual stimulus. Auditory stimuli can also alter the perception of visual scenes in which there is an underlying reality but which are ambiguous due to noise in sensory estimators, as

demonstrated in the 'illusory flash' or 'fission' effect (Shams et al., 2000). This effect occurs when a single visual flash is presented together with two or more auditory beeps, leading to the perception of more than one flash. A related illusion is the 'fusion' effect (Andersen et al., 2004) in which fewer beeps than flashes are presented, and the observer perceives fewer flashes as a result. Both of these effects have been widely studied in adults; in addition to psychophysical evidence for perceptual integration (e.g., McCormick & Mamassian, 2008), neuro-imaging studies have shown that auditory stimuli used in these tasks can modulate activity in primary visual cortex (e.g., Shams, Iwaki, Chawla & Bhattacharya, 2005a; Shams, Kamitani, Thompson & Shimojo, 2001; Watkins, Shams, Tanaka, Haynes & Rees, 2006; Watkins, Shams, Josephs & Rees, 2007).

Very few studies have examined the developmental timecourse of fission and fusion effects. Tremblay et al. (2007) compared these effects across three age groups: 5-9, 10-14 and 15-19 years old. They found that no difference existed in the number of either fission or fusion illusions between the different age groups. They suggest that this is due to the audio-visual integration mechanisms necessary for the task developing very early: before 5 years old. However, it is not clear that 15-19 year olds would experience the same audio-visual percepts as mature adults: no adult control group was reported in this study. The fusion effect appears from their data to be weaker (although not significantly) in 15-19 year olds than in the younger groups, possibly suggesting continuing maturation of audio-visual integration. The size of fission or fusion effects does not directly allow evaluation of audio-visual integration mechanisms. To determine whether an optimal Bayesian integration strategy is used (Section 1.3) unimodal percepts must be measured in order to generate predictions of responses to both cue conflict and congruent bimodal trials.

Tremblay et al. (2007) compared the strength of the fission and fusion illusions between age groups but did not use unimodal response data to predict conflict trial responses so could not conclude whether an optimal integration strategy was used by any of the groups. Despite this they conclude that audio-visual mechanisms are mature by the age of their youngest participants: 5 years old. In a more recent study of fission and fusion illusions in children, Innes-Brown et al. (2011) compared a group of 8-17 year olds with adults using the fission/fusion task. In contrast with Tremblay et al. (2007), they claim that audio-visual integration is immature in 8-17 year olds, being greater and less selective than in adults, with children experiencing more fission illusions. However, these claims may be unjustified: although they find a larger fission effect in children, it could be that this is as a result of an optimal integration strategy but with greater uncertainty in visual estimates. The higher uncertainty (variance) in children's visual estimates could result in the auditory estimate being weighted more heavily and hence dominating. Since there are no unimodal auditory trials reported, it is not possible to compare the reliability of each modality's estimate or to test for optimal integration.

The contradictory conclusions of Tremblay et al. (2007) and Innes-Brown et al. (2011) stem from a common weakness in their approaches: they do not measure visual and auditory unimodal estimates and so are unable to distinguish whether differences in children's and adults' experiences of the fission/fusion illusions result from different integration strategies or a common strategy with different relative cue reliabilities.

A number of studies have suggested that sensory integration develops relatively late, becoming statistically optimal between 8 and 12 years old (Gori et al., 2008; Nardini et al., 2008, 2010, see also Section 1.5). Nardini et al. (2008) found that, in contrast with adults, children below 8 years old did not integrate visual cues with vestibular and proprioceptive cues to benefit from a reduction in uncertainty. Nardini et al. (2010) showed that two visual cues, texture and disparity, are not optimally integrated in 6 year olds, resulting in greater uncertainty in slant estimates but improved ability to discriminate between the two cues as compared with adults. The uncertainty reduction due to sensory integration was found to be fully developed by 12 years old. Visual-haptic integration is also sub-optimal before 8 years old, with one sense dominating the other even if the dominant sense is not the most reliable (Gori et al., 2008). Gori et al. found that between the ages of 8 and 10 years old this visual-haptic integration becomes statistically optimal. Late development of sensory integration (both within and between senses) has been posited to allow access to individual sensory estimates, and so enable detection of conflicts between these estimates; this in turn may provide the error signals required to drive recalibration as the body develops (Gori et al., 2008; Nardini et al., 2010). Each of these studies involves either comparison of multiple stimuli (Nardini et al., 2010; Gori et al., 2008) and/or working memory demands (Gori et al., 2008; Nardini et al., 2008). In young children, the noise added by these additional task demands may overwhelm any small decreases in bimodal variance (compared with unimodal variances) such that even if they used an optimal integration strategy, it might not be detected (Nardini et al., in press).

Whilst it is helpful to consider the audio-visual integration abilities of children over the age of 8 (as Innes-Brown et al. did), given the evidence from these studies that sensory cue combination matures between approximately 8 and 12 years old, it would also be useful to consider the abilities of children younger than 8 years old. Whereas previous studies tested the size of the fission/fusion effect in childhood, to test whether cue combination is optimal it is necessary to measure unimodal cue estimates to generate predictions for bimodal conditions. The study presented here compares a group of 5-7 year olds with adults on both the bounce/stream and fission/fusion tasks. These tasks offer two tests of integration abilities: the bounce/stream task considers whether sound can affect the percept of a visual stimulus; the fission/fusion task is a stronger test which, through modelling responses, allows quantitative predictions regarding how sound affects the percept of a visual stimulus. Neither of these tasks has memory demands beyond reporting the percept on each trial.

4.2 Experiment 1: Bounce/Stream

4.2.1 Method

4.2.1.1 Participants

Observers were students at the University of Southampton (adults) and pupils at a local primary school (children). None of the participants were experienced psychophysical observers and all were naïve to the purposes of the experiment. All observers had normal or corrected-to-normal vision and normal hearing and gave informed consent before the experiment, which was approved by the local ethics committee. In addition to the child's consent, parents of child observers gave informed consent for their child to participate.

Twenty-seven adults participated in exchange for course credit. The mean age was 20.8 years (standard deviation = 2.5 years). Fifty-six children participated in this study; the mean age was 79.9 months (6.7 years, standard deviation = 7.0 months). Most (52) of these children also participated in the second experiment.

4.2.1.2 Apparatus

Stimuli were generated and displayed using Matlab 2009b and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner, Brainard & Pelli, 2007). Visual stimuli were presented on either a MacBook Pro or MacBook running OS X version 10.6.6. The viewing distance was 60cm, in a room with normal levels of ambient light. Auditory stimuli were presented via headphones.

4.2.1.3 Stimuli

Visual stimuli were white discs subtending an angle of 1° with a luminance of 15.0cd/m^2 on a mid-grey background. The discs appeared in the top corners of the screen and travelled diagonally across the centre at a speed of 6° per second, forming an X shape (see Figure 4.1)). These stimuli produce one of two perceptual interpretations: on reaching the screen's centre, the discs either appear to (i) continue their path, 'streaming' past each other, or (ii) 'bounce' off each other. To allow participants to understand and identify the two percepts, practice trials used two discs of different colour and luminance (dark blue and light pink) to create unambiguous 'bounce' or 'stream' trials. On experimental (ambiguous) trials the discs were both white. However, pilot data showed that on ambiguous trials with identical discs, children perceived bouncing on nearly all trials (μ=81%; 8 out of 12 participants responded bouncing on more than 90% of trials). To allow us to explore the effects of

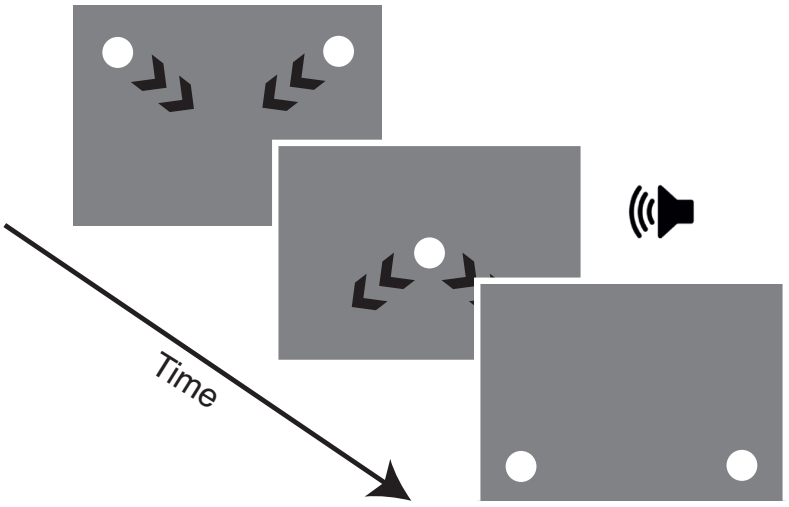

Figure 4.1: **Schematic of stimuli for 'bounce/stream' experiment**

auditory stimulus components, we manipulated the experimental stimuli to bias the children's percepts toward streaming: the luminance of one disc was decreased slightly (14.1cd/m^2) and its size was increased (to subtend an angle of 1.04°). This manipulation was extremely difficult to detect, even for the experimenters who were aware of the differences. Observers were not informed of the bias, however it did produce more balanced 'bounce' and 'stream' responses. For adults, the two discs were identical on experimental trials. Trials were either silent or accompanied by a 17ms beep of 1500Hz (56dB, chosen to be clearly audible but not uncomfortably loud), 150ms before coincidence, at coincidence or 150ms after coincidence.

4.2.1.4 Procedure

Eight unambiguous practice trials (four bounce, four stream, half with a beep at coincidence, half without) were used to check that the participant understood the task. Adults completed 80 experimental trials (20 repetitions of each of 4 sound conditions: absent, pre, at and post coincidence). Children completed 20 trials (5 repetitions x 4 sound conditions). In both cases, observers made a two-alternative forced choice (2AFC) between whether the discs bounced or streamed past each other by selecting an icon representing the movement. Adults sat alone to complete the study whereas a researcher sat with each child. There were up to four children participating in the room at one time.

	Absent ($\mu = 0.40$)	**Pre** ($\mu = 0.60$)	**Coincident** ($\mu = 0.67$)
Pre ($\mu = 0.60$)	$p < 0.001^\star$	-	-
Coincident ($\mu = 0.67$)	$p < 0.001^\star$	$p = 0.042^\star$	-
Post($\mu = 0.60$)	$p < 0.001^\star$	$p = 1$	$p = 0.046^\star$

Table 4.1: **Comparisons of all participants' 'bounce' responses by timing of auditory stimulus**
Pairwise comparisons of the proportion of 'bounce' responses in each of the auditory timing conditions. Means are given in the column and row headings and all p-values are Bonferroni corrected; * indicates significant differences ($p < 0.05$). Timing of auditory stimulus is relative to the coincidence of visual discs.

4.2.2 Results

For each auditory condition and for each participant the proportion of trials was calculated in which the discs were perceived to have bounced rather than streamed. An ANOVA was used to explore the effects of age group (adults, children) and auditory condition (sound absent, pre-, at- and post-coincidence)

There was no main effect of age group; our stimulus manipulation was successful in biasing child observers towards a streaming response, producing the same overall proportion of 'bounce' responses for adults and children (adults vs children, $F(1, 81) = 0.21$, $p > 0.05$).[1] Observers were sensitive to both the presence and relative timing of auditory stimuli (main effect of auditory condition: $F(3, 243) = 25.19$, $p < 0.001$, G-G corrections, $\epsilon = 0.80$). Participants experienced significantly more bouncing percepts when the auditory stimulus was present than absent and significantly more bouncing percepts when the auditory stimulus was at the discs' coincidence than when it was offset in either direction (see Table 4.1 for details).

There was, however, an interaction between auditory condition and age group ($F(3, 243) = 4.61$, $p = 0.007$, G-G correction $\epsilon = 0.80$). For children, 'bounce' responses were significantly more prevalent when the beep was present than absent (pairwise comparisons shown in Figure 4.2). However, bounce responses were not significantly different between the three beep present conditions (pre, coincident, and post).

In contrast, adults were sensitive not only to the presence or absence of the auditory stimulus, but also to its timing relative to the visual stimulus. The proportion of bouncing responses was highest when the auditory stimulus was coincident with the discs intersecting and lowest when the auditory stimulus was absent. When the auditory stimulus was presented 150ms pre- or post- the intersection of the discs, the proportion of bouncing trials was significantly higher than in the absent condition and

[1] However, as children got older the proportion of trials which they perceived as 'bounce' trials reduced (correlation of age in months vs. mean 'bounce' responses across all auditory conditions, $r = -0.34$, $p = 0.01$).

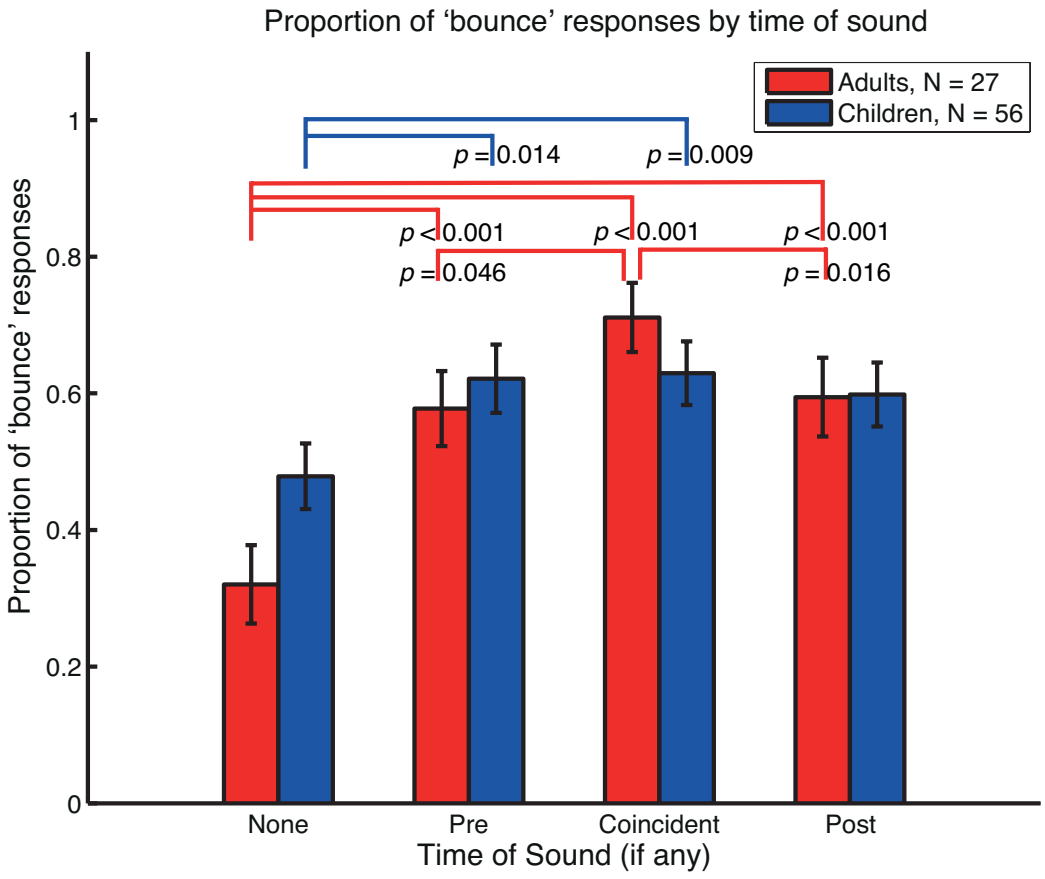

Figure 4.2: **Mean proportion of bounce responses at each timing interval.** Adult data is shown in red; child data is shown in blue. Significant pairwise comparisons (Bonferroni corrected) are shown. In addition to these, there was a marginally significant difference in the children's data between the proportion of trials perceived as bouncing for the beep absent condition and the post coincidence beep condition ($p = 0.073$). Error bars represent ± 1 SEM.

significantly lower than in the coincident condition (see Figure 4.2 for details of pairwise comparisons). This adult pattern is consistent with previous results (Sekuler et al., 1997).

To explore more fully the different patterns observed in children and adults the possible causes of the interaction were separated into two factors: firstly, the presence or absence of the auditory stimulus could affect the percepts of children and adults differently; secondly, the two groups may be differently sensitive to the relative timing of the auditory stimulus compared with the visual stimulus. To test for the former, the difference between sound present (coincident) and sound absent conditions were compared in adults and children. The presence of the auditory stimulus promoted bouncing percepts (main effect of sound present: $F(1, 81) = 47.48$, $p < 0.001$). There was no effect of age although there was a significant interaction: adults experienced a larger increase than children in the proportion of bouncing percepts when the sound

	Absent	**Coincident**
Adults	0.32	0.71
Children	0.48	0.63

Table 4.2: **Mean proportion of 'bounce' responses by presence of auditory stimulus in adults and children**

	Coincident	**Offset**
Adults	0.71	0.59
Children	0.63	0.61

Table 4.3: **Mean proportion of 'bounce' responses by temporal alignment of auditory and visual stimuli in adults and children**

was present compared with when it was absent ($F(1, 81) = 9.31$, $p = 0.003$, means shown in Table 4.2).

Before testing for the possibility that the interaction may have been partially caused by differences in responses with different relative timing, the 'pre' and 'post' conditions were tested to check for significant differences. Responses to pre and post conditions were not significantly different for either adults or children and so an average was taken and named the 'offset' condition. The offset condition was then compared with the coincident condition in adults and children. Participants were sensitive to the relative timing of the auditory stimulus compared with the visual stimulus (main effect of relative timing: $F(1, 81) = 10.24$, $p = 0.002$). There was no effect of age but there was a significant interaction between age and relative timing: adults appeared to be more sensitive to the relative timing than children, experiencing a greater reduction in bouncing percepts than children when the auditory stimulus was offset rather than aligned with the discs' visual coincidence ($F(1, 81) = 5.43$, $p = 0.022$, means shown in Table 4.3).

Taken together, these results suggest that whilst both adults and children are sensitive to the presence of an auditory stimulus when interpreting a visual scene, the presence of that auditory stimulus has a greater effect on visual percepts for adults than it does for children. In addition, children are less sensitive to the relative timing of the auditory and visual stimuli than adults: they may integrate over a wider time window than adults.

Given the greater sensitivity of adults to the relative timing of the auditory stimulus, it is possible that the older children we tested would show some increase in sensitivity compared with the younger children. However, a correlation between age (in months) and sensitivity to relative timing (proportion of 'bounce' responses in coincident condition - proportion of bounce responses in offset condition) showed no increase in sensitivity with age (at least not within $\pm$150ms, $r = 0.16$, $p = 0.24$).

4.3 Experiment 2: Fission/Fusion

4.3.1 Method

4.3.1.1 Participants

Observers were students at the University of Southampton (adults) and pupils at a local primary school (children). None of the participants were experienced psychophysical observers and all were naïve to the purposes of the experiment. All observers had normal or corrected-to-normal vision and normal hearing and gave informed consent before the experiment, which was approved by the local ethics committee. In addition to the child's consent, parents of child observers gave informed consent for their child to participate.

In the second experiment (flash/beep), forty adults participated in exchange for course credit. The mean age was 21.0 years (standard deviation = 2.3 years). Sixty children participated in this study; the mean age was 80.3 months (6.7 years, standard deviation = 6.8 months).

4.3.1.2 Apparatus

Apparatus were as described in Section 4.2.1.2.

4.3.1.3 Stimuli

Visual stimuli were white discs that subtended an angle of 2° with a luminance of 23.8cd/m^2 on a black background, horizontally displaced by 5° either to the left or right of fixation. For each flash, the disc was presented for 1 frame (16.7 ms) and off for 4 frames (adults) or 9 frames (children) (see Figure 4.3). The auditory stimulus was either absent or concurrent with the flashes. One, two or three beeps (7 ms long, 440Hz, 60dB) were played, starting simultaneously with the onset of the first flash; flash and beep frequencies were matched such that subsequent beeps were aligned to subsequent flashes when present.

4.3.1.4 Procedure

A black screen was presented with a white fixation cross in the centre. The participant clicked on this cross to start the trial, ensuring central fixation. On each trial the white disc flashed on either once, twice or thrice and participants reported the number of perceived flashes (one, two or three). Adults completed 8 repetitions of each trial type

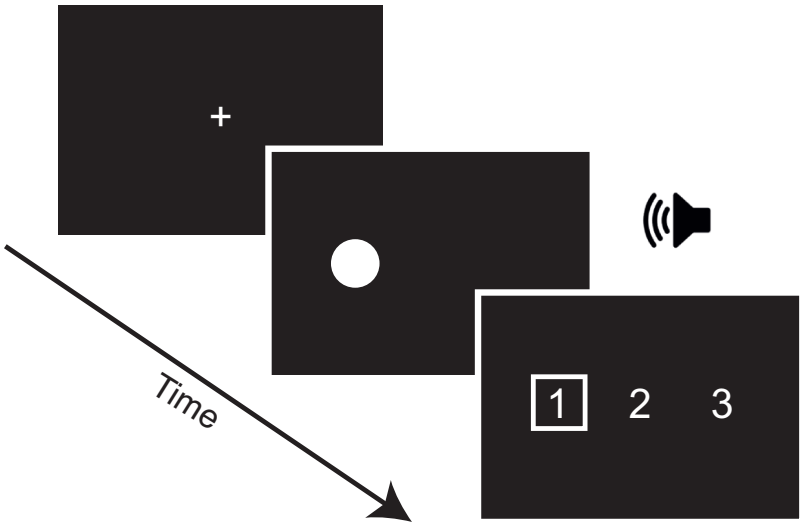

Figure 4.3: **Schematic of stimuli for 'flash/beep' experiment**

		n_{beeps}			
		0	**1**	**2**	**3**
$n_{flashes}$	**0**	-	✓	✓	✓
	1	✓	✓	✓	✓
	2	✓	✓	✓	-
	3	✓	✓	-	✓

Table 4.4: **Combinations of stimuli used in flash/beep trials**
Ticks mark the combinations of number of flashes and beeps used; F2B3 and F3B2 were omitted to reduce the number of trials that children completed.

(half presented to the left and half presented to the right of fixation), see Table 4.4 for detail as to which trial types there were. After the audio-visual trials, there were 8 repetitions of auditory-only trials (also shown in Table 4.4); participants responded to say how many beeps they heard. The procedure was identical for children and adults except that children completed only 2 repetitions. Again, adults completed the experiment alone whereas children sat with a researcher.

4.3.2 Results

4.3.2.1 Analysis

The strength of the fusion effect was defined as the difference in mean response between the two and three flash visual-only trials and the corresponding two or three flash trials with a single beep (i.e., $F2B0 - F2B1$ and $F3B0 - F3B1$). If the auditory stimulus has no effect, i.e., there is no fusion effect, this difference will be zero. The larger the difference in these two responses (more positive), the greater the fusion

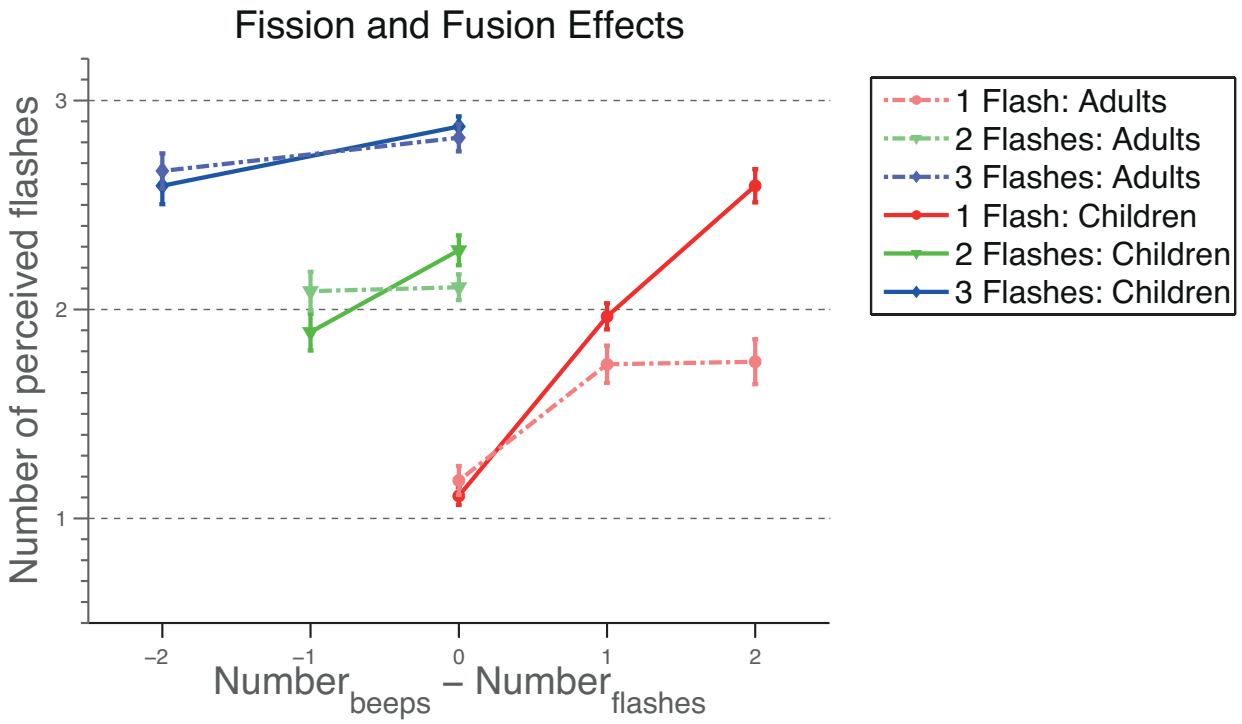

Figure 4.4: **Mean number of flashes perceived as a function of audio-visual discrepancy.**
Adult data (N = 40) is shown with dashed lines, child data (N=60) is shown with solid lines. The deviation of data lines from horizontal shows the effect of the number of beeps on the perceived number of flashes. The fusion effect (where present) is shown on the left half of the graph (number of beeps < number of flashes). The fission effect is shown on the right half of the graph (number of beeps > number of flashes). Error bars represent ±1 SEM.

effect. Similarly, the strength of the fission effect was defined as the difference in mean response between the single flash trials with either 2 or 3 beeps and the visual-only single flash trials (i.e., $F1B2 - F1B0$ and $F1B3 - F1B0$). Again, if the auditory stimulus has no effect, i.e., there is no fission effect, this difference will be zero. The larger the difference between these responses (more positive), the greater the fission effect. The fusion and fission effects are illustrated by the means shown in Figure 4.4.

4.3.2.2 Fusion Effects

The fusion effect was tested by comparing the fusion strength (as defined in Section 4.3.2.1) across three conditions: baseline (unimodal, fusion strength = $F2B0$ - $F2B0$ = $F3B0$ - $F3B0$ = 0); small audio-visual discrepancy (fusion strength = $F2B0$ - $F2B1$); and large audio-visual discrepancy (fusion strength = $F3B0$ - $F3B1$). Participants, across age groups, experienced significant fusion effects (main effect of audio-visual discrepancy on fusion strength: $F(2, 196) = 16.07$, $p < 0.001$). Compared

with unimodal stimuli, participants experienced a significant fusion effect when presented with two flashes and one beep but not when there were three flashes and one beep ($\mu_{F2B0-F2B1} = 0.30$, $p < 0.001$; $\mu_{F3B0-F3B1} = 0.13$, $p = n.s.$, Bonferroni corrected comparisons). There was significantly more fusion in the two flash, one beep condition than the three flash, one beep condition ($p = 0.008$, Bonferroni corrected). Children experienced a significantly greater fusion effect than adults (main effect of age group on fusion strength: $F(1, 98) = 10.13$, $p = 0.002$; $\mu_{adults} = 0.05$, $\mu_{children} = 0.24$). There was also an interaction between audio-visual discrepancy and age group ($F(2, 196) = 6.47$, $p = 0.002$): adults showed only a marginally significant fusion effect when two flashes were presented with a single beep ($\mu_{F2B0-F2B1} = 0.11$, $p = 0.085$, Bonferroni corrected) and no fusion effect when three flashes were presented with a single beep ($\mu_{F3B0-F3B1} = 0.04$, $p = n.s.$, Bonferroni corrected); children, by contrast, experienced significant fusion effects in both cases ($\mu_{F2B0-F2B1} = 0.49$, $p < 0.001$; $\mu_{F3B0-F3B1} = 0.22$, $p = 0.046$, Bonferroni corrected). However, the fusion effect was stronger when the audio-visual discrepancy was smaller ($p = 0.006$, Bonferroni corrected).

4.3.2.3 Fission Effects

The fission effect was tested by comparing the fission strength (as defined in Section 4.3.2.1) across three conditions: baseline (unimodal, fission strength = $F1B0$ - $F1B0 = 0$); small audio-visual discrepancy (fission strength = $F1B2$ - $F1B0$); and large audio-visual discrepancy (fission strength = $F1B3$ - $F1B0$). Participants, across age groups, experienced significant fission effects (main effect of audio-visual discrepancy on fission strength: G-G correction for non-sphericity, $\epsilon = 0.89$; $F(2, 196) = 129.24$, $p < 0.001$). Compared with unimodal stimuli, participants experienced a significant fission effect when presented with either one flash and two beeps or one flash and three beeps ($\mu_{F1B2-F1B0} = 0.50$, $p < 0.001$; $\mu_{F1B3-F1B0} = 0.82$, $p < 0.001$, Bonferroni corrected comparisons). There was significantly more fission when the audio-visual discrepancy was larger ($p < 0.001$, Bonferroni corrected). Children experienced a significantly greater fission effect than adults (main effect of age group on fission strength: $F(1, 98) = 12.10$, $p = 0.001$; $\mu_{adults} = 0.33$, $\mu_{children} = 0.56$). There was also an interaction between audio-visual discrepancy and age group ($F(2, 196) = 25.33$, $p < 0.001$): adults showed significant fission effects when a single flash was presented with either two or three beeps ($\mu_{F1B2-F1B0} = 0.48$, $p < 0.001$; $\mu_{F1B3-F1B0} = 0.49$, $p < 0.001$, Bonferroni corrected) but there was no significant difference between these two cases; children also experienced significant fission effects in both cases ($\mu_{F1B2-F1B0} = 0.53$, $p < 0.001$; $\mu_{F1B3-F1B0} = 1.15$, $p < 0.001$, Bonferroni corrected), however, unlike adults the fission effect was stronger when audio-visual discrepancy was greater ($p < 0.001$, Bonferroni corrected).

4.3.2.4 Modelling

In agreement with previous studies (Tremblay et al., 2007; Innes-Brown et al., 2011), the preceding results show that the fission illusion is stronger in children than in adults; furthermore, we find this also to be the case for the fusion illusion. However, it is not clear that this is the result of children using a different cue combination strategy to adults: an alternative explanation may be that they are using the same strategy as adults, but that their unimodal estimates have different relative reliabilities compared with adults and hence are weighting the two cues differently to adults. It is clear from the response means that both adults and children are using the information from both modalities to perceive the number of flashes on any given trial, but the means on their own do not provide sufficient information to determine cue combination strategy. There are at least two possible mechanisms by which these results could be obtained: both optimal integration and 'switching' between the two cue estimates could result in the observed means. However, the two strategies would result in different response variances for different trial types. For non-conflict (congruent) trials (e.g., F1B1, F2B2), optimal partial integration should result in a response variance lower than that measured in visual-only trials; a switching strategy could result in either a lower or a higher response variance than that measured in visual-only trials, depending on whether the visual and auditory unimodal estimates were closely aligned or not. For conflict trials (e.g., F1B3, F2B1) optimal partial integration predicts the same decrease in response variance as for congruent trials; however, a switching model predicts that response variance will increase in cue-conflict trials since the means of the two estimates will not be aligned. Here we compare these two models of cue combination, using measurements of unimodal response variance to generate predictions for cross-modal trials; these predictions are then compared to the measured response variance.

Unimodal variances for 1, 2 and 3 flashes/beeps were calculated separately for each participant before averaging across participants; this was done to remove the effect of different response biases which could have artificially inflated an overall (cross-participant) measure of response variance. Since there were no significant differences between the variances for 1, 2 or 3 events in either modality, the unimodal variances were also averaged across number of events to give a single variance for each modality (σ_v^2 and σ_a^2 for visual and auditory variance respectively). These were used as inputs to the models described below to generate predictions of variance in bimodal conditions. The empirically measured response variance in each of the bimodal conditions was also calculated separately for each participant before averaging across participants.

Another step common to both models is to calculate the strength of the effect of auditory stimuli on visual estimates using the method described by Bresciani et al.

(2006). A linear regression was calculated between response error and the difference between the number of beeps and flashes; the gradient of the regression line provides a measure of the auditory influence on vision. Response error was calculated by subtracting the mean response (across all participants) for the unimodal visual trials from the mean response to equivalent bimodal trials (i.e., trials with the same number of flashes but varying numbers of beeps). If the gradient of the regression line is zero that may be interpreted as the auditory stimulus having no influence on the visual percept: vision and audition are not combined. If, on the other hand, the gradient of the line is one, this may be interpreted as the visual percept being entirely dominated by the auditory percept. A gradient over 0.5 would mean that the auditory estimate was weighted more than the visual estimate, a gradient less than 0.5 would mean that that the visual estimate was weighted more than the auditory estimate. Both children and adults gave more weight to the visual estimate than the auditory estimate, although children gave more weight to audition than adults did: for adults, the gradient $\Delta_{v|a} = 0.17$ ($R^2 = 0.64$); for children, $\Delta_{v|a} = 0.38$ ($R^2 = 0.72$).

4.3.2.5 Coupling Prior

The first model uses a coupling prior (Ernst, 2006; Bresciani et al., 2006); this represents the joint distribution of the two world properties being perceived (in this case the number of flashes and number of beeps). If the beeps and flashes are believed to have the same cause, the coupling prior takes the form of the identity line $n_{beeps} = n_{flashes}$. If the beeps and flashes are thought to be completely unrelated, the coupling prior is completely uniform such that any combination of n_{beeps} and $n_{flashes}$ is equally probable. In between these two extremes, the coupling prior is aligned with the identity line, but with a Gaussian spread. The prior then encodes the belief that the two cues are somewhat likely to have the same cause, but may not always. The probability that the two world properties have the same cause is represented by the variance of the Gaussian spread. The two extremes described above can be modelled in the same way, but with zero variance (full coupling) or infinite variance (no coupling) respectively. The optimal Bayesian integration described in Section 1.3.2 is consistent with Bayesian integration using a coupling prior that has zero variance: the two signals are completely integrated and the weights are based solely on the reliability of the two cues. The coupling prior is multiplied with the joint likelihood distribution for the two cue estimates to obtain the posterior (illustrated in Figure 4.5). Subsequently a decision rule is applied to the posterior to obtain a bimodal estimate; here we use MAP estimation (as defined in Section 1.2.2).

To calculate the variance of the coupling prior for each group of participants (adults and children) we used the equations set out in Bresciani et al. (2006): the relative

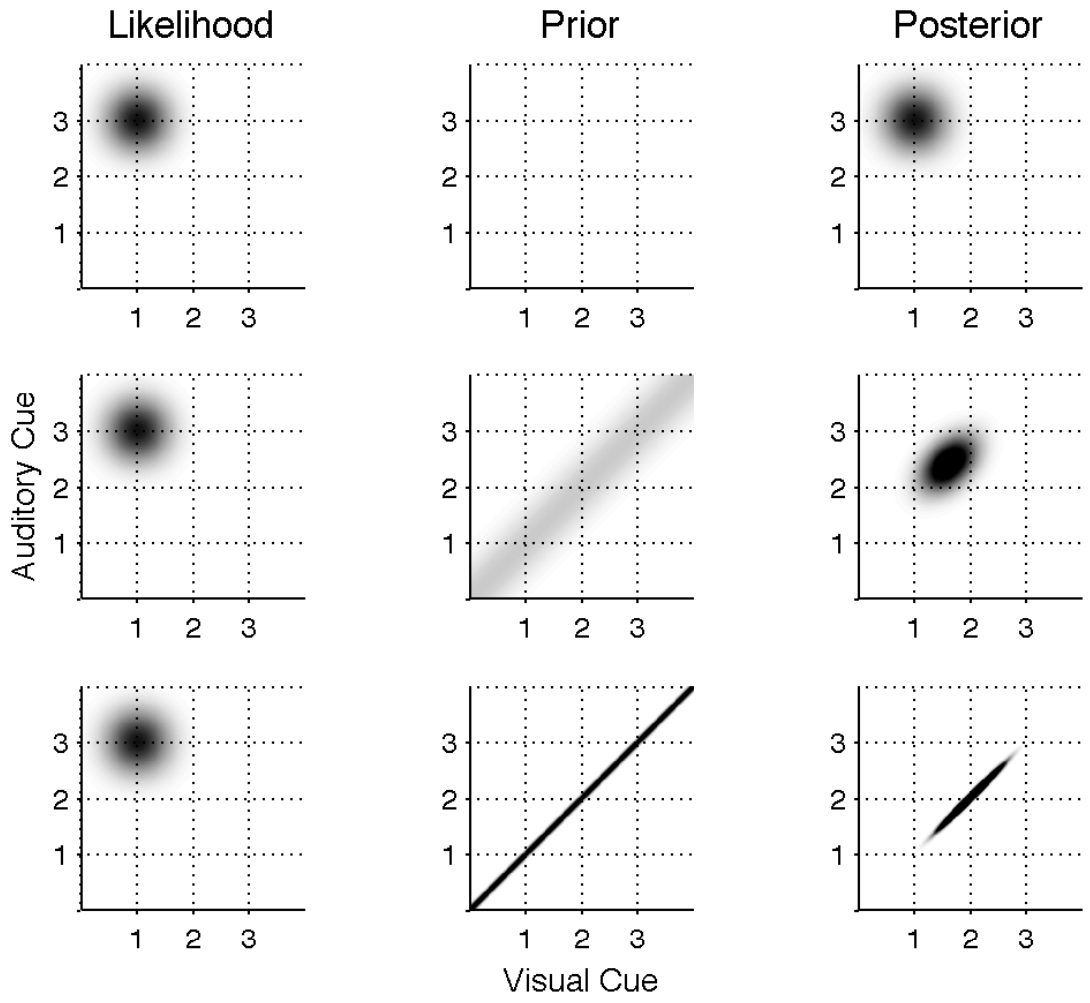

Figure 4.5: **Examples of the effect of coupling prior variance on posterior distribution**
Each row shows a 2D likelihood, coupling prior and resultant posterior probability distributions: darker areas show higher probability. The likelihoods in these examples have equal variance in the visual and auditory cue estimate (equally reliable cues). The visual cue in this example is 1 flash and the auditory cue is 3 beeps and so the combined likelihood is centred on 1 flash/3 beeps. The coupling prior varies in each row: in the first row it is small and uniform; this is equivalent to a belief that the two cues are independent and so any combination is equally likely (no coupling). The resulting posterior distribution is positioned identically to the likelihood. The second row shows a Gaussian coupling prior with a small variance: this is equivalent to the belief that two events often, but not always, have the same underlying cause. The final estimate lies between the likelihood and the prior. The third row shows an example of full coupling: the coupling prior is a delta function, this corresponds to the belief that the two events are completely determined by one another. The final estimate lies along the line of the prior since all other combinations are impossible.

influence of audio on vision,

$$\alpha = \arctan \frac{\sigma_v^2}{\sigma_a^2} = \arctan \frac{\Delta_{v|a}}{\Delta_{a|v}}, \tag{4.1}$$

where $\Delta_{a|v}$ is the change in auditory percepts due to visual stimuli (this was not measured in this experiment, but rather calculated using Equation 4.1). Then degree of coupling,

$$C = \Delta_{v|a} + \Delta_{a|v}, \tag{4.2}$$

and

$$C = \frac{\sigma^2_{likelihood}(\alpha)}{\sigma^2_{likelihood}(\alpha) + \sigma^2_{prior}(\alpha)}, \tag{4.3}$$

with

$$\sigma^2_{likelihood}(\alpha) = \sigma_v^2 \cos^2(\alpha) + \sigma_a^2 \sin^2(\alpha), \tag{4.4}$$

and

$$\sigma_{prior}(\alpha) = \sigma_p \cos(|\alpha - 45°|). \tag{4.5}$$

Solving for σ_p^2 gives the coupling prior variance; the prior itself is calculated as a Gaussian distribution 'extruded' along the identity line. In this model a lower variance indicates stronger integration as the expectation that $n_{flashes} = n_{beeps}$ is higher. Likewise, a higher variance indicates weaker integration and a lower expectation that $n_{flashes} = n_{beeps}$. This coupling prior is independent of current stimulus values.

For each bimodal condition we use participants' unimodal data to approximate the likelihood as a two-dimensional Gaussian distribution centred on (μ_v, μ_a) (the mean responses for unimodal trials with the corresponding number of flashes and beeps respectively) with variances (σ_v^2, σ_a^2) (the unimodal response variances as calculated in Section 4.3.2.4). The posterior probability distribution for the current condition is calculated by multiplying the likelihood by the prior and normalising (see Figure 4.6); the variance in the visual axis of the resulting distribution is the predicted response variance for that condition (see Figure 4.7).

4.3.2.6 Switching Model

The second model is a cue switching model; such models have previously been applied to other cue combination tasks (e.g., Landy & Kojima, 2001). Here we model the responses for the number of flashes as being drawn variously from the visual-only response distribution and the auditory-only response distribution. The weight given to auditory information (the change in visual estimates due to the audio influence) was calculated using the slope of the same regression line as in the coupling prior model ($\Delta_{v|a}$). The proportion of times in which responses were drawn from the auditory-only likelihood distribution (L_a) was $\Delta_{v|a}$. The proportion of times that responses were

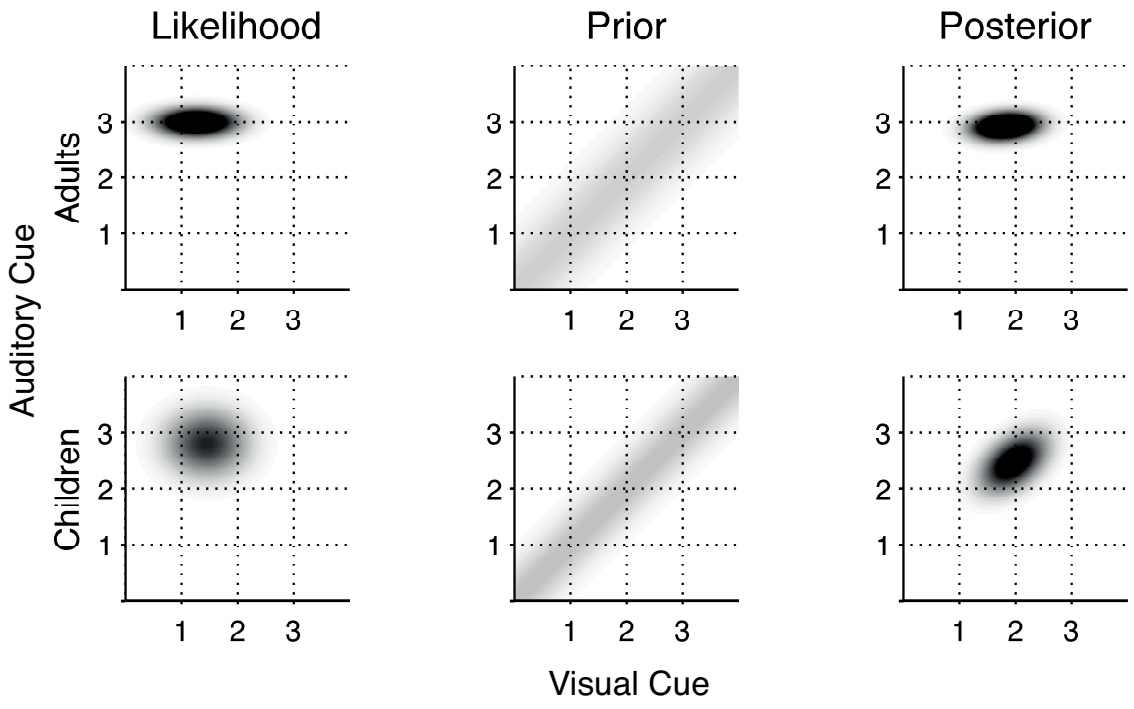

Figure 4.6: **Plots of likelihood, coupling prior and posterior distributions in adults and children**
As in Figure 4.5, each row shows a 2D likelihood, a coupling prior and the resultant posterior probability distributions; again, darker areas have higher probability. Just as in Figure 4.5 this example shows distributions for 1 flash and 3 beeps; the combined likelihood is centred on the mean unimodal responses to 1 flash and 3 beeps respectively. The first row shows the adults' data: the reliability of the auditory cue is higher than that of the visual cue ($\sigma_{auditory} = 0.15$, $\sigma_{visual} = 0.42$, respectively) which is why the likelihood is relatively compressed in the auditory axis. The coupling prior is a Gaussian with $\sigma_p = 0.42$. The second row shows the children's data: the auditory and visual cue reliabilities are more similar ($\sigma_{auditory} = 0.36$, $\sigma_{visual} = 0.45$, respectively) and so the likelihood is less compressed in the auditory axis than the adults' likelihood. The coupling prior is a Gaussian with $\sigma_p = 0.31$, this is narrower than the adult's coupling prior, suggesting a stronger belief that the events in the two modalities have a common cause.

taken from the visual-only likelihood distribution (L_v) was $1 - \Delta_{v|a}$. The visual and the auditory likelihood probability density functions are one-dimensional Gaussian distributions with means μ_v and μ_a, and variances σ_v^2 and σ_a^2 for vision and audition respectively. The combined likelihood (L_{av}) for each bimodal condition is then calculated as:

$$L_{av}(n_{flashes}, n_{beeps}) = (1 - \Delta_{v|a})L_v(n_{flashes}) + \Delta_{v|a}L_a(n_{beeps}). \tag{4.6}$$

The predicted response variance for the switching model is the variance of L_{av}, which varies with $n_{flashes}$ and n_{beeps} (see Figure 4.7).

Participant Group	Coupling	Switching
Adults	0.0046	0.0368
Children	0.0089	0.1037

Table 4.5: **Mean Square Error of response variance predictions for the Coupling Prior and Switching models**

4.3.2.7 Comparison of Models

To compare the two models of cue combination with the empirical data, the mean square error (MSE) was calculated across conditions between the coupling and empirical response variances, and between the switching and empirical response variances, i.e.:

$$MSE_{model} = \frac{\sum_{i=1}^{n_{conditions}} (\sigma^2_{model}(i) - \sigma^2_{empirical}(i))^2}{n_{conditions}} \tag{4.7}$$

Table 4.5 shows the results of this analysis; for both adults and children, the MSE is considerably smaller for the coupling model than the switching model, suggesting that the former is better able to predict response variances in bimodal trials.

A qualitative prediction of the coupling prior model is that response variance would be the same in congruent and conflict bimodal trials. However, despite the coupling prior model being a better fit to the data than the switching model, there was a significant increase in response variance in conflict trials compared with congruent trials (adults: $t(39) = 5.85$, $p < 0.001$, $\mu_{congruent} = 0.12$, $\mu_{conflict} = 0.21$; children: $t(59) = 3.08$, $p = 0.003$, $\mu_{congruent} = 0.13$, $\mu_{conflict} = 0.22$). This suggests that although the coupling prior model provides a better fit than the switching model, it is not able to explain fully the observed data.

4.4 Discussion

The results from both experiments show that both adults and children combine information from visual and auditory cues when generating visual percepts.

In Experiment 1 we replicated the bounce/stream effect previously found in adults (Sekuler et al., 1997) and extended this to show that children also experience increased 'bounce' percepts in the presence of an auditory stimulus. Previously, Scheier et al. (2003) found that 6 month old infants experience the bounce/stream illusion; by contrast, the children in our study experienced bouncing percepts in the vast majority of trials. This necessitated the addition of a bias towards streaming in the visual stimuli to enable measurement of the effect of auditory stimuli on visual perception. Even with the addition of a streaming bias, the proportion of 'bouncing' responses

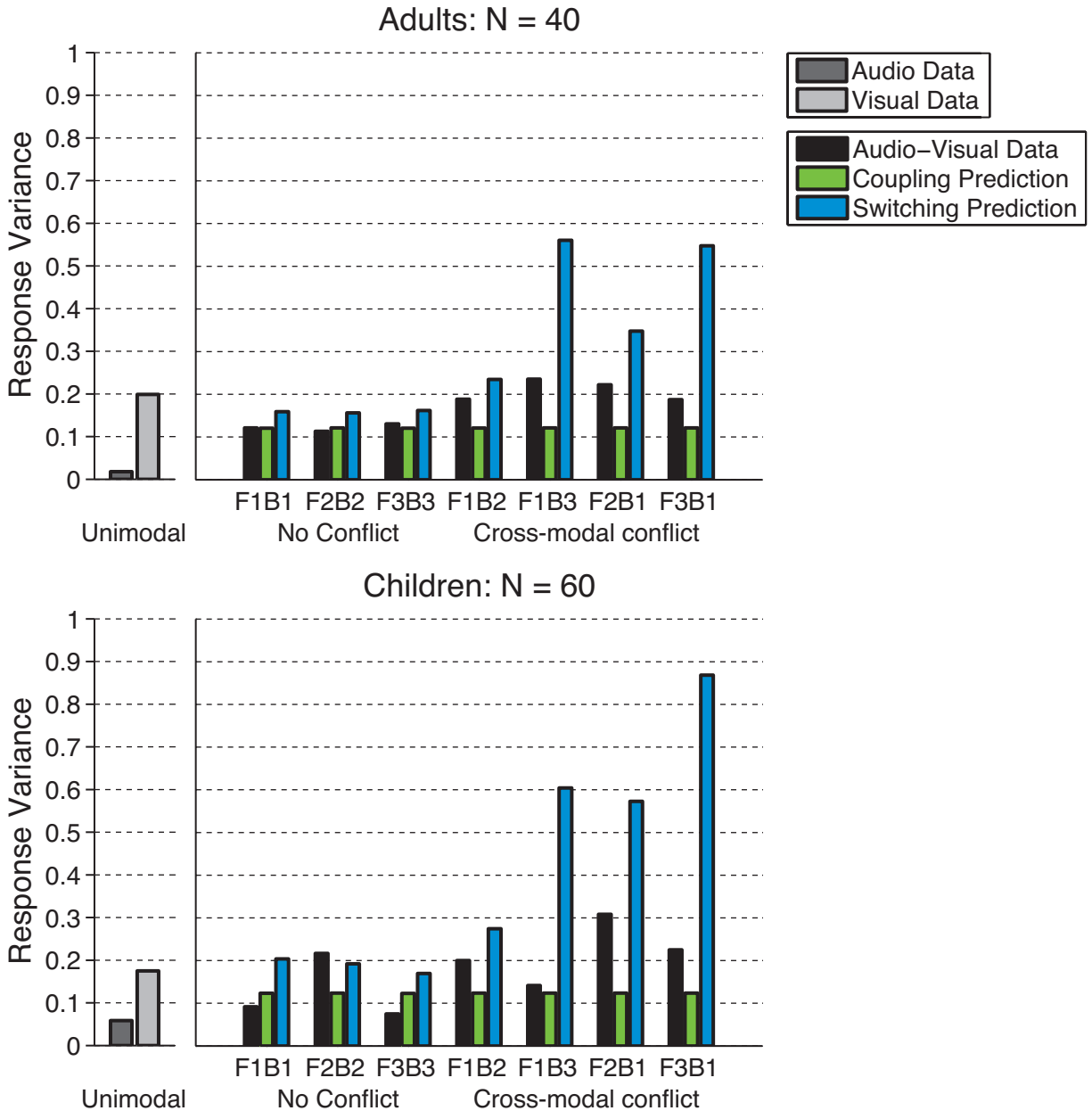

Figure 4.7: **Variance for data and model predictions**
The audio and visual variance are for unimodal trials. The audio-visual variance and predicted variances are for bimodal trials - as indicated on the x-axis.

reduced with increasing age. Given the difficulty in interpreting looking-time data, it may be that the infants in Scheier's study, like our children, predominantly experienced a bouncing percept and were instead responding to temporal changes in the auditory stimuli rather than perceiving the bounce/stream illusion.

The children in our study also demonstrated less sensitivity to the temporal offset between visual and auditory stimuli than adults. This suggests that they may have less certainty in their temporal judgements and so are less able to determine whether two events were synchronous. A consequence of this increased uncertainty might be the need to employ a longer integration window to benefit from cue combination. To quantify the difference in integration window between adults and children, it would be necessary to present visual and auditory stimuli with a broader range of temporal offsets and to measure the accuracy of their temporal judgements.

There is another explanation of the bounce/stream effect which argues the effect is (at least partially) due to attentional disruption: in adults bouncing percepts increased both when there were distractors present and when there was an additional discrimination task (Watanabe & Shimojo, 1998). They suggest that attention may enable the perception of continuous motion. In this case the increased bouncing percepts in childhood may be as a result of immature attentional processes.

Experiment 2 provided a stronger test of audio-visual cue integration. In common with previous studies we found that adults experience the fission illusion (e.g., Shams et al., 2000). We found only marginally significant evidence that adults experience the fusion illusion reported by others (e.g., Andersen et al., 2004). We demonstrated that children experienced the fission illusion to a greater extent than adults (as in Innes-Brown et al., 2011) and furthermore that they also experience a stronger fusion illusion than adults. Innes-Brown et al. (2011) also reported increased fusion illusions in children (although not significantly), and the data reported by Tremblay et al. (2007) also appears to show a trend for more frequent fusion illusions in younger compared with older children, although again not significantly. Innes-Brown et al. (2011) concluded that:

> *"These results show that the mechanisms that integrate auditory with visual integration, giving rise to the flash-beep illusion, do not follow a linear developmental trend with age in this group of normally developing children."*

On the contrary, we have shown here that differences in frequency of illusion between adults and children are not necessarily the result of immature integration mechanisms, but rather may be caused by optimal integration of less reliable unimodal estimates in children. We compared a partial integration (coupling prior) model (Bresciani et al., 2006) and a switching model with the empirical data, and found the former to generate better predictions of response variance in both adults and children. The means were

qualitatively consistent with both the switching and the partial integration model, but quantitative predictions of the means cannot be made since they were used to predict the response variance.

These results are consistent with the visual-haptic integration reported by Bresciani et al. (2006); our data also supports the findings of Shams et al. (2005b) that the fission and fusion illusions result from optimal cue integration. Both their model and the coupling prior approach used here incorporate a finite probability that the two cues have separate causes. As such, rather than fully integrating the two cue estimates as in the standard Bayesian model, they are partially integrated such that estimates in one modality have a lesser effect on estimates in the other modality than would be predicted from their relative reliabilities alone.

Previous studies (e.g., Gori et al., 2008; Nardini et al., 2008, 2010) found that optimal sensory integration develops relatively late in childhood, maturing between approximately 8 and 12 years old. In contrast, we have shown that in the case of audio-visual cues, optimal integration strategies may already be in place in our participants (5-7 years old, mean age 6.7 years old). The late development of cue integration strategies has been thought to arise from a need to continue to calibrate cues through childhood, for example recalibrating haptics as the body grows (Gori et al., 2008). However, optimal cue integration in adults does not preclude continuing cue calibration: Adams et al. (2004) found that haptics can recalibrate the light-from-above prior and Adams et al. (2010) found that binocular disparity could also recalibrate the light-from-above prior.

Although the coupling prior integration model generated better quantitative variance predictions than the switching model, neither model exactly matched the empirical data. Indeed, the switching model better fitted the pattern of response variance from a qualitative point of view: response variance increased in conflict trials compared with congruent trials. There are several reasons why the models may not match the observed data: firstly the response method truncated the range of responses such that observers could never respond more than three or less than one. This prevented response variances from being symmetric in the one and three event conditions, which could have affected both the predictions and empirical responses in the bimodal conditions. The unimodal response variances used to generate the predictions were also likely to have been underestimated; due to the very small number of trials per participant (adults: 8, children: 2) and the relatively high reliability of the auditory stimuli, a number of participants exhibited accuracy rates of 100% resulting in a variance estimate of zero. In future studies it would be helpful to increase the number of trial repetitions and expand the range of allowed responses to enable better measurement of variances. However, it might not be possible to increase the number of trials for children by a great deal as they have a limited concentration span. Decreasing the reliability of the auditory stimuli might also mitigate against the

problem of measuring zero variance. However, the stimuli for the current study were chosen to allow close comparison with previous studies; this would be more difficult with modified stimuli.

Finally, the coupling prior model could be tested further by collecting auditory responses to the same audio-visual stimuli (i.e., reporting the number of beeps heard). The posterior probability distribution generated by the coupling prior model can be used to make predictions of both visual and auditory responses to bimodal stimuli (by finding the MAP estimate in each dimension); if the two modalities have similar reliabilities and integration is not complete (i.e., the coupling prior has non-zero variance) then visual responses to bimodal stimuli would be predicted to lie closer to the corresponding unimodal visual estimate then the unimodal auditory estimate, and vice versa. Measuring auditory responses would also give direct access to a measurement of $\Delta_{a|v}$ (effect of visual stimuli on auditory responses), such that the coupling prior variance could be calculated with fewer assumptions, as in Bresciani et al. (2006). This measurement would also benefit from the two modalities having similar reliabilities: if, as in the current study, the unimodal auditory estimates have a much lower variance than the unimodal visual estimates then the effect of visual stimuli on auditory estimates ($\Delta_{a|v}$) is very small and hence hard to detect. Therefore, it would be helpful in future studies to decrease the reliability of the auditory stimuli such that the two modalities have similar reliabilities.

Looking beyond temporal audio-visual cue integration, the ventriloquist effect is a well-established example of optimal integration of spatial cues in adulthood; studying this effect in children of a similar age to the present study would provide further evidence as to the development of children's cue integration strategies.

In summary, both experiments suggest that audio-visual integration is well developed in 5 to 7 year olds; children employ an optimal partial integration strategy, just like adults, but experience audio-visual illusions differently to adults due to differences in unimodal cue reliabilities and differences in the variance of their coupling priors. Neither experiment provides evidence to suggest that adults integrate auditory and visual information in a qualitatively different way to children as had been suggested previously for visual-haptic, visual-vestibular and visual-visual cue integration (Gori et al., 2008; Nardini et al., 2008, 2010).

Chapter 5

Learning different light prior distributions for different contexts

Kerrigan, I. S. & Adams, W. J. (under review at Cognition) Learning different light prior distributions for different contexts.

Experimental design, data collection, analysis and write-up were completed by Iona Kerrigan under the supervision of Wendy Adams. Data from six participants in the mixed condition were submitted for Iona Kerrigan's MSc dissertation at the University of Southampton; it has been reanalysed, together with data from the other twenty participants, for this chapter.

5.1 Abstract

The pattern of shading across an image can provide a rich sense of object shape. Our ability to use shading information is remarkable given the infinite possible combinations of illumination, shape and reflectance that could have produced any given image. Illumination can change dramatically across environments (e.g., indoor vs. outdoor) and times of day (e.g., midday vs. sunset). Here I show that people can learn to associate particular illumination conditions with particular contexts, to aid shape-from-shading. Following a few hours of visual-haptic training, observers modified their shape estimates according to the illumination expected in the prevailing context. Our observers learned that red lighting was roughly overhead (consistent with their previous assumption of lighting direction), whereas green lighting was shifted by 10°. Learning was more efficient when training for the two contexts (red or green light) was mixed rather than sequentially blocked.

5.2 Introduction

Humans cope with reddish illumination at sunset or flickering coloured lights at the disco - managing to decompose shading patterns into reflectance and shape variations - but how? Our impressively robust ability to estimate our surroundings, given complex and ambiguous retinal input relies heavily on prior knowledge - we bias perceptual estimates toward the most likely scenes. For example, we bias estimates of illumination direction toward overhead (e.g., Kleffner & Ramachandran, 1992; Adams, 2007) and estimates of surface shape toward convexity (Langer & Bülthoff, 2001; Adams & Mamassian, 2004) in alignment with the statistics of our environment (Potetz & Lee, 2003). The perceptual assumptions or 'priors' used for the tasks of material perception (Chapter 2 and Chapter 3) are not well researched; however, it is well known that light priors facilitate the notoriously under-constrained problem of recovering shape-from-shading (e.g., Kleffner & Ramachandran, 1992; Adams, 2007). This chapter builds on our knowledge of light priors to investigate whether the adult human visual system can learn and selectively invoke multiple context-specific priors.

For optimal performance, humans should (i) respond to long-term changes in scene statistics by updating their priors and (ii) select the correct prior for a given context. We know that humans do the former: in contrast with chickens (Hershberger, 1970), human observers change their light prior in response to appropriate haptic (Adams et al., 2004) or visual feedback (Adams et al., 2010). Here I ask whether humans also do the latter: can we learn different prior assumptions for different contexts? There is no clear consensus: although Adams et al. (2004) found that a modified light-prior generalised to novel stimuli, Adams et al. (2010) noted that modified light-priors were retained for several weeks beyond training, after observers had returned to their normal environment, in which lighting was presumably, on average, overhead. This latter finding suggests that observers learnt separate, context-dependent light priors, with the experimental set-up acting as a contextual cue.

Here I ask whether humans can learn two light priors, each invoked by a different illumination colour. To induce colour-dependent learning, visual-haptic feedback was modulated by the simulated illumination colour: when scenes were illuminated by red light, feedback was consistent with the observer's baseline light prior distribution. In contrast, under green illumination, feedback was consistent with a new lighting distribution.

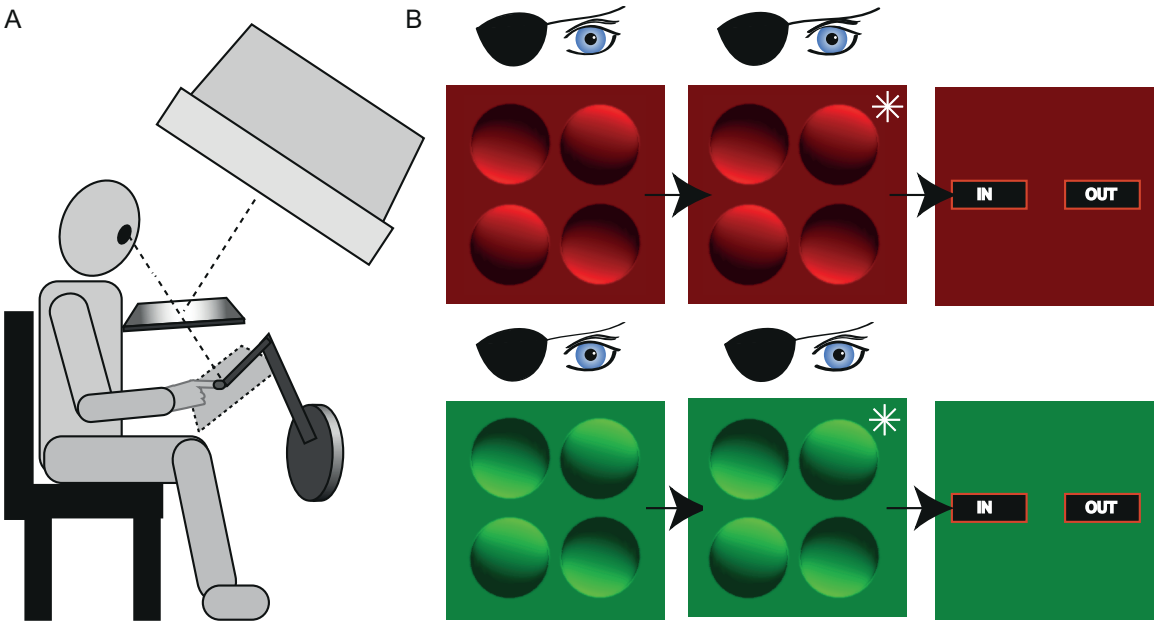

Figure 5.1: **Apparatus & visual test trials.**
(A) The visual-haptic experimental set-up. (B) Examples of visual-only test trials: the simulated lighting is either red (upper row) or green (lower row). Observers briefly viewed the four shaded discs before indicating whether the cued object was concave or convex (in or out).

5.3 Methods

5.3.1 Apparatus & Stimuli

Observers simultaneously viewed and felt virtual objects (see Figure 5.1A). Haptic scenes were presented via a 'thimble gimbal' attached to a force-feedback device (Ghost libraries, PHANToM, SensAble Technologies). Visual stimuli (Figures 5.1B & 5.2C - 5.2F), generated using OpenGL, were presented via a front-silvered mirror. Their perceived location (at a visual distance of 56cm) matched the location of the haptic stimuli, giving the impression of a single visual-haptic scene. A headrest and bite-bar maintained head position and an eye-patch eliminated binocular depth cues. The room was completely dark, other than the light emitted by the visual display.

5.3.2 Visual Test Trials

Pre- and post-training trials contained solely visual (no haptic) information. Observers viewed four shaded discs, each subtending 5.6° and offset from the screen's centre by 5.3° (see Figure 5.1). Each disc was consistent with a hemisphere squashed in depth by a factor of 2, illuminated by a distant light source. The slant of the light source (the angle between the lighting vector and the screen normal) was 68.2°. The light source

tilt (the angle between the projected lighting vector and the vertical axis in the plane of the screen, θ) varied across trials. This illumination tilt, with object shape (convex vs. concave) determined the shading orientation of each disc. Within each trial, one, two or three discs had a shading gradient direction of θ and the remaining disc(s) had a shading gradient of $\theta + 180°$, such that observers generally perceived both convex and concave objects to be present. The simulated scene was white, with either a red or green simulated light source although stimuli were equally consistent with red and green scenes illuminated by white light.

Observers judged the shape (concave vs. convex) of one object (cued by a star). The observer's light prior was estimated from the set of 288 visual trials (24 equally spaced θ values x 2 colours x 6 repetitions), lasting approximately 10-15 minutes (see Figure 5.2A).

5.3.3 Training Trials

Visual-haptic training was similar to that used previously (e.g., Adams et al., 2004, see Figures 5.2C - 5.2F). Observers viewed four shaded discs (as in test trials), but also explored the scene haptically by running a finger (in a thimble gimbal) over the simulated objects. This haptic information disambiguated each object's shape, and thus also the lighting direction. However, the relationship between shading orientation and haptic shape depended on colour (see Figure 5.2B). On 'red' trials, stimuli were consistent with the observer's baseline light prior; haptic shape matched the observer's pre-training shape responses. On 'green' trials, however, the lighting direction was drawn from a range shifted by $\pm 30°$ relative to the observer's baseline prior (13 observers were assigned a $+30°$ shift, 13 a $-30°$ shift). Thus, on 'green' trials, some objects previously perceived as convex now felt concave, and vice versa.

After haptically exploring the scene for a minimum of 7s, including 'touching' all four objects, the observer pressed a button to continue. One of the objects then appeared visually (without haptics) in the centre of the screen for 1s and the observer judged its shape (convex/concave). By subsequently viewing and touching the object, observers gained feedback on their response. Each training set comprised 224 visual-haptic training trials (48 equally spaced θ values x 2 colours x 2 repetitions + 2 extra repetitions of 8 θ values within conflict regions x 2 colours), lasting approximately 60-90 minutes. There is some evidence that people and animals learn to discriminate between two contingencies more quickly when trials are intermixed than when they are blocked (e.g., Mitchell, Nash & Hall, 2008; Honey, Bateson & Horn, 1994). To identify whether a similar advantage is observable for this context-dependent learning task, I assigned observers to either an (i) mixed or (ii) blocked variant. In the first variant, red and green trials were randomly mixed throughout test and training. In the second, colour was fixed within blocks of 24 trials.

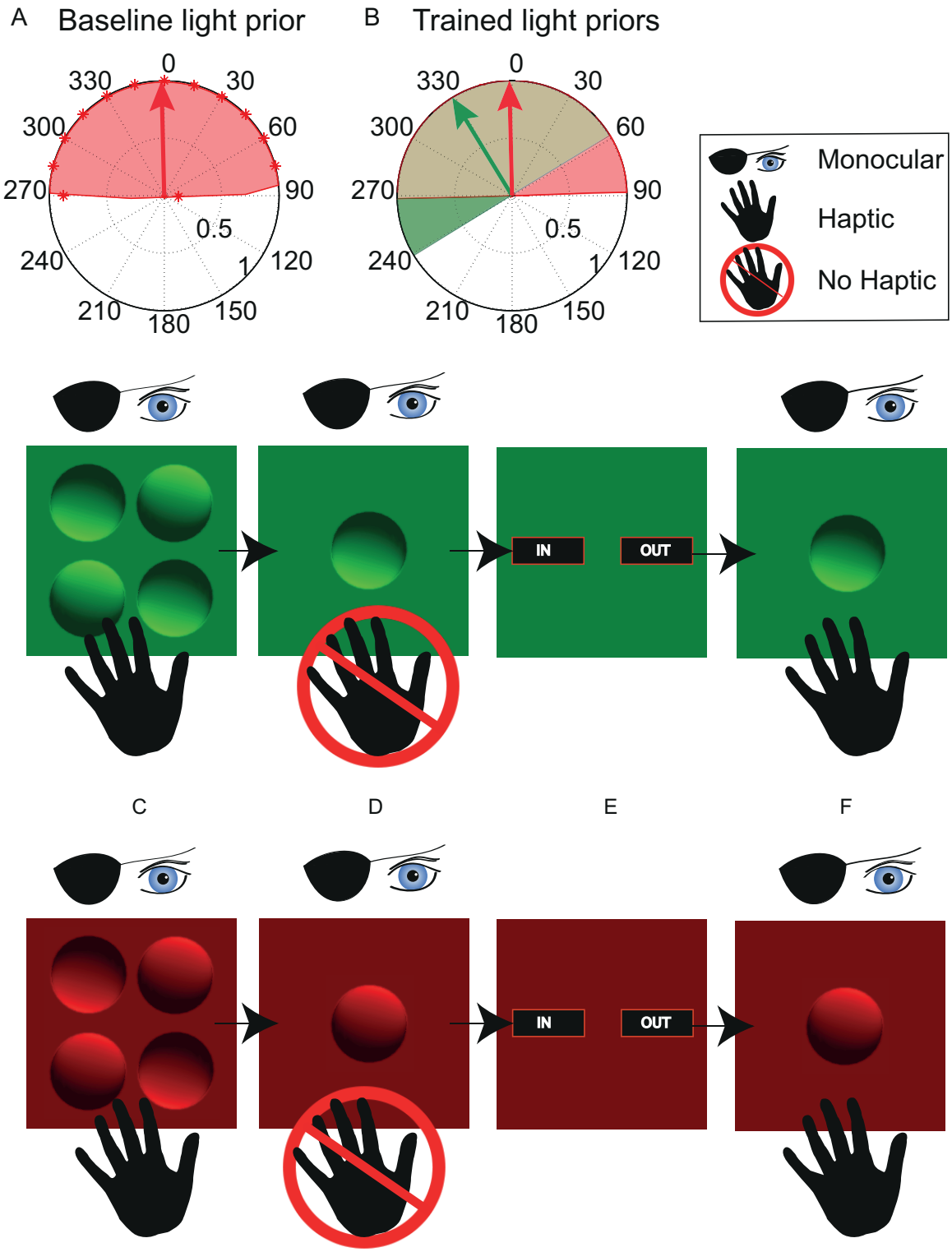

Figure 5.2: **Visual-haptic training trials.**
See caption on next page.

Figure 5.2: (A) Proportion of responses that were convex, as a function of shading orientation, for one observer. The shaded region thus also reflects their baseline light prior distribution, with the mean of the light prior given by the red arrow. (B) Two trained light priors. The red region indicates the range of stimulus shading orientations that 'felt' convex during red visual-haptic training trials - it matches the range of orientations perceived as convex at baseline. The green region indicates the orientations that 'felt' convex during green visual-haptic trials. (C-F) Schematic representation of a training trial: (C) Observers explored the scene both haptically and visually and then (D) viewed a single disc for 1s before (E) judging its shape. (F) Viewing and touching the single stimulus provided feedback.

5.3.4 Procedure

On day 1, each observer completed a set of visual-only trials, followed (after a short break) by a train-test session (one set of visual-haptic training trials, one set of visual test trials). On day 2, they completed two train-test sessions, separated by at least one hour.

5.3.5 Participants

Twenty-six naïve observers completed the experiment (mixed variant: 10 participants; blocked variant: 16 participants). All had normal or corrected-to-normal visual acuity and normal colour vision. Participants gave informed written consent and the local ethics committee approved the study.

5.3.6 Possible Outcomes

What might observers learn from the visual-haptic training? What colour-dependent or colour-independent changes in shape perception might be seen? First, observers might show no learning: if colour were ignored as a 'nuisance' variable, haptic feedback would appear noisy and inconsistent and thus might be discounted. Second, observers might ignore colour, but still modify their behaviour to reflect the aggregate of all feedback. Their light priors would thus move toward the average of the two trained lighting distributions, irrespective of stimulus colour. Finally, observers may learn (consciously or unconsciously) that particular colours are associated with particular lighting distributions. This would allow them to apply different prior distributions over lighting direction in different colour contexts. This context-specific learning would result, post-training, in different measured light priors for different coloured test stimuli - the same shading orientation would induce different perceived shapes under different illumination colours.

5.4 Results

Light priors were estimated by fitting a simple Bayesian model to each observer's test and training data (see Adams et al. (2010); essentially, the peak of the light prior is given by the peak of the 'convex' responses). I first checked whether the perceived shape of red and green stimuli differed prior to training. Three observers were excluded (one from the mixed condition, two from the blocked condition) as their red and green baseline priors differed ($ps = 0.015$; 0.015; and 0.003, from bootstrapping). A single baseline light prior was estimated for each remaining observer, using their combined red and green pre-training data ($\mu = -9.61°$, $\sigma = 13.61°$, across observers). All subsequent data were separated by colour to estimate colour-specific light priors.

Results are shown in Figure 5.3. Training had a significant effect on shape perception ($F(3, 63) = 6.61$, $p = 0.003$, $\epsilon = 0.67$, G-G correction for non-sphericity, from 3 factor ANOVA (amount of training, training type and illumination colour), partial $\eta^2 = 0.24$). As training progressed, observers' light prior distributions moved toward the trained lighting direction. Significant learning occurred after two sets of training (normalised baseline light prior = 0°, vs. mean penultimate and final light priors $\mu = 7.71°$ and $\mu = 9.45°$, $p = 0.012$ and $p = 0.042$ respectively, from Bonferroni corrected comparisons).

Importantly, observers did show some context dependent learning: light priors were shifted significantly further from baseline when measured with green than with red test stimuli ($F(1, 21) = 5.28$, $p = 0.032$, partial $\eta^2 = 0.20$). However, learning was not entirely context dependent: significant changes in light prior were observed for both illumination contexts by the end of training (green: $\mu = 9.77°$, red: $\mu = 6.61°$, t-tests against zero: $p = 0.011$ and $p = 0.045$, respectively).

In line with previous work (e.g., Mitchell et al., 2008; Honey et al., 1994), those trained in the intermixed condition showed significantly more learning than those in the blocked condition (15.25° vs. 3.64° in the final test session, significant main effect of training type (blocked vs. intermixed): $F(1, 21) = 4.31$, $p = 0.05$, partial $\eta^2 = 0.30$). No interactions were significant.

5.5 Discussion

I show that the visual system is able to learn and implement separate light priors for different contexts. After training, the perceived shape of ambiguous shaded objects was modulated by illumination colour. Observers were able to switch between two different, colour-contingent light priors on a trial-by-trial basis. However, this colour-contingent behaviour was implemented unconsciously - at debrief, observers

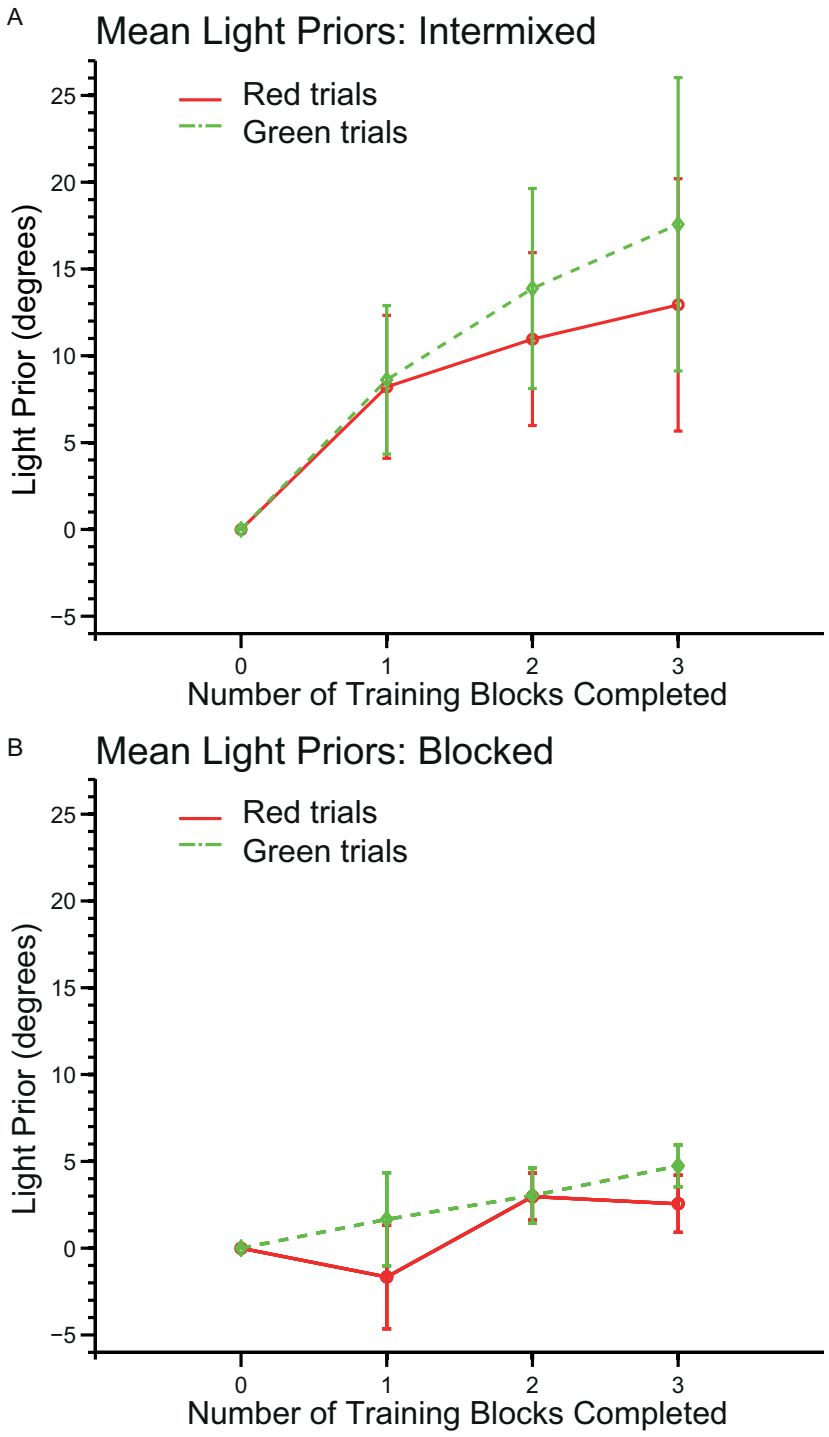

Figure 5.3: **Light priors before and after training**
(A) in the intermixed condition and (B) blocked condition. To allow meaningful comparisons across observers, each observer's data were normalised by his or her baseline light prior, and light priors for observers who trained with a -30° shift were multiplied by -1.

were unaware of any differences between red and green stimuli, consistent with recent evidence that observers can learn cue relationships of which they are unaware (Di Luca et al., 2010). Context-dependent use of priors has clear benefits: implementing a light prior that accurately reflects the lighting statistics of the current context will lead to more accurate shape judgements. Interestingly, however, the observed learning was not entirely context-specific; observers did not fully differentiate on the basis of colour but instead showed a combination of colour-specific and colour-independent learning. This may reflect robustness to temporary, perhaps spurious changes in cue contingencies; previous experience suggests that illumination direction and colour are not strongly correlated. In other words, this learning response may not reflect a limitation in our ability to learn, but rather an optimal strategy given the likelihood of such changes. This robustness has been modelled using Kalman filters where weight is given to both historical and current input (e.g., Burge, Ernst & Banks, 2008). A similar logic applies to the incomplete learning found here (final average 'green' light-prior 9.77°, vs. trained light-prior of 30°) and in previous studies with comparable training, where observers learned only a third (11°) of the trained light prior shift (Adams et al., 2004).

Other research has investigated the priors over aspect ratio that contribute to orientation estimation (an elliptical retinal image may be perceived as a slanted circle). Knill (2007) demonstrated that observers use visual-haptic feedback to modify their prior on the aspect ratio of ellipses. Similarly to Adams et al. (2010), Knill noted that observers' priors did not readapt on exposure to the normal environment, suggesting that learning was specific to the laboratory context. Later, Seydell et al. (2010) demonstrated that observers can learn separate aspect ratio priors for different shapes (diamonds vs. ellipses) but appear unable to learn aspect ratio priors conditioned on object colour. The authors suggest that colour cannot be used to modulate observers' shape priors because colour is deemed to be unrelated to aspect ratio by the visual system: there is no ecological reason to link colour with aspect ratio. This study shows that colour can act as a contextual cue; it may be that a relationship between illumination colour and illumination direction is deemed more plausible by the visual system.

A similar argument is presented by Michel & Jacobs (2007). They suggest that learning will be relatively easy when an existing relationship is modified (parameter learning). In contrast, learning a new relationship (structural learning) will be difficult or impossible: they tested whether observers could learn an association between illumination direction and stimulus depth, but concluded that they could not. Under this framework, observers learned the distinction between colour contexts because colour and illumination direction are sometimes related in the real world: the relationship has ecological validity. However, it remains unclear whether, given enough training, observers would be able to learn two different light priors for contextual cues that are ecologically unrelated to illumination (e.g., object shape or texture). Ernst

(2007) demonstrated that people can learn to associate low-level cues (luminance and stiffness) that have previously been unrelated. This suggests ecological validity is not a necessary condition for learning.

In broad agreement with Michel & Jacobs (2007), Backus and colleagues have shown that some cue associations are easier to learn than others (for a review see Backus, 2011). They asked which novel cues may be 'recruited' such that they influence the interpretation of an ambiguous rotating structure from motion (SFM) stimulus. Haijang et al. (2006) and Jain et al. (2010) found that some cues (location, motion direction) were recruited as contextual cues that modulated SFM perception. However, they found that other cues (e.g., auditory cues and extrinsic visual cues) were not recruited to disambiguate the SFM stimulus (but see also Backus, Jain & Fuller, 2011). I suggest that when a pair of cues has previously been unrelated in the environment, the visual system should represent this information; strong evidence that particular signals are unrelated allows the visual system to avoid learning new, spurious relationships. In contrast, learning will be faster when the visual system holds little information about whether or not two signals are correlated, or data that they are sometimes correlated. In this way, modification of cue relationships may be better characterised as a continuum rather than by dichotomies such as parameter vs. structural learning.

In summary, I show that colour can be learned and used as a cue to context: observers are able to selectively invoke different light priors in different contexts, allowing accurate recovery of shape from shading in their current environment. This use of context-dependent priors will assist the visual system as it moves between lighting environments, particularly when direct information about the prevailing illumination conditions is ambiguous.

Chapter 6

Discussion

6.1 Motivation for thesis

Perception is an ill-posed, under-constrained problem (Poggio & Torre, 1984) and so sensory systems must make use of probabilistic relationships between sensory data and world properties. Over the past 20 years or so, the Bayesian framework has proved to be a flexible method for implementing probabilistic models of many different perceptual cues, both individually and in combination. It has been shown to be a strong predictor of human perception, able to account for many different experimental effects (e.g., Yuille & Bülthoff, 1996; Ernst & Banks, 2002; Knill & Saunders, 2003; Bresciani et al., 2006; Körding & Wolpert, 2004; Weiss, Simoncelli & Adelson, 2002).

Within the Bayesian framework, perceptual estimates are generated from the combination of one or more likelihoods (the probability of the current sensory data given each world state) with one or more priors, which instantiate the assumptions necessary to interpret the sensory data (the probability of each world state independent of the current sensory data). Where sensory cues are independent of each other they can each be modelled by an individual likelihood; where they are not independent they may be modelled by a single joint (multi-dimensional) likelihood. The product of likelihood and prior probability distributions is the posterior, which represents the probability of each world state given the current sensory data. The posterior in combination with some decision rule gives a perceptual estimate of the world property (Maloney, 2002b). Two commonly used decision rules are *Maximum a Posteriori* or MAP estimation: the peak of the posterior distribution is chosen as the perceptual estimate; and maximum likelihood estimation (MLE) which is the same as MAP estimation if a uniform prior is assumed.

In our natural environment there are usually myriad sensory cues available that provide some information about various properties such as object shape and material.

The redundant information they contain can improve the precision of perceptual estimates, if the cues are used together (e.g., Ernst & Banks, 2002; Alais & Burr, 2004). There are various models of cue combination (described in further detail in Section 1.3.1), each of which may be instantiated using the Bayesian framework. Strong fusion (Clark & Yuille, 1990) describes the case where cues may interact with each other in arbitrarily complex ways. From a Bayesian perspective this is modelled as a single likelihood distribution and a single prior (which may be uniform) (e.g., Nakayama & Shimojo, 1992). These probability distributions cannot be factorised into the constituent cues' likelihoods and priors. At the other end of the spectrum, weak fusion (Clark & Yuille, 1990) considers all cues to be separable and, where they are independent from one another, to be linearly combined. This has the advantage that cues can be studied in isolation and then used to generate predictions of perceptual performance when multiple cues are available. Modified weak fusion is a compromise between the two extremes in which cues may interact with one another for the purposes of cue 'promotion' but are otherwise linearly combined (Landy et al., 1995). Most studies of cue combination follow a modified weak fusion approach but assume (either implicitly or explicitly) that cue promotion has already happened (e.g., Adams et al., 2004; Maloney, 2002a). One general method to implement weak or modified weak fusion, using the Bayesian framework, is to multiply all the individual cue likelihoods and priors together to give the joint posterior distribution (Yuille & Bülthoff, 1996). The estimate can then be selected from the posterior distribution in the same way as for an individual cue.

Perceptual cues are combined both within (e.g., Knill & Saunders, 2003; Hillis et al., 2002, 2004) and across (e.g., Ernst & Banks, 2002; Alais & Burr, 2004; Hillis et al., 2002) modalities to improve the precision and accuracy of perception. Most cue combination studies have focused on geometric properties such as shape (Ernst & Banks, 2002; Helbig & Ernst, 2007b; Wijntjes et al., 2009) or location (Alais & Burr, 2004), or temporal properties such as the number of events (Shams et al., 2005b; Bresciani et al., 2006). At least as important for effective interaction with the environment is knowing the material properties of objects and surfaces (Adelson, 2001). For example, when walking on a surface it is useful to adapt one's gait depending on the properties of the surface: whether hard or soft and whether slippery or rough. Similarly, an estimate of the strength and rigidity of an object may give some information as to its suitability for use as a tool. There are several possible reasons for the relative paucity of research into material perception: Adelson (2001) suggests a linguistic bias in our tendency to discuss 'things' rather than 'stuff' but there are also practical issues in experimental design. Limitations in the computational capability of graphical hardware and software have meant that it is only relatively recently that rendering any realistic materials has been possible. There are still some limitations, for example in rendering the effects of inter-reflections between multiple objects and in rendering the reflective properties of more complex materials such as

human skin which exhibits subsurface scattering of light (Jimenez, Sundstedt & Gutierrez, 2009). Similar constraints exist in the rendering of haptic material properties. For example, the PHANToM technology used in the experiments in Chapter 2 and Chapter 5 can replicate proprioceptive forces (friction and compliance) but not fine grained textures or other tactile and thermal properties. Some of these limitations could be avoided by using real objects within an augmented reality set-up similar to that used by Di Luca, Knörlein, Ernst & Harders (2011).

To further our understanding of the perception of material properties, Chapter 2 and Chapter 3 investigated gloss perception. There are many cues to material properties; Chapter 2 specifically asked whether the haptic cues of friction and compliance could affect perceived gloss. It also considered whether the binocular disparity of highlights affected gloss perception. Highlight disparity contributes not only to gloss perception but also to shape perception (Blake & Bülthoff, 1990, 1991): Chapter 3 explored the relationship between highlight disparity and perceived shape and gloss, specifically considering whether an accurate model of highlight geometry is implemented to constrain both shape and gloss percepts.

Although the combination of spatial and temporal cues is well established in adulthood, less is known about how and when these capabilities develop. There has been mixed evidence as to the age at which cue combination strategies mature, with some studies suggesting that the use of multimodal cues is possible in infancy (Scheier et al., 2003) and others suggesting that adult strategies do not develop until 10 years old, or older (Gori et al., 2008; Nardini et al., 2008; Innes-Brown et al., 2011). Chapter 4 used two experiments to test audio-visual integration capabilities in children and compared their performance to that of adults, within a single quantitative framework. The first experiment addressed a relatively broad question: do auditory stimuli affect visual percepts in the target age group? The second experiment compared two different models of behaviour: (i) a coupling prior model based on the Bayesian framework and (ii) a switching model, to test whether differences in performance between adults and children are due to differences in the precision of individual estimators and/or the priors used, or whether they are implementing qualitatively different cue combination strategies.

Cue combination is one way to improve the reliability of perceptual estimates; another way in which multiple cues can be used to improve perception is in the recalibration of likelihoods and priors. Both likelihoods and priors can be recalibrated over time in order to reflect more accurately the current environmental statistics (e.g., Adams et al., 2001, 2004) but there is less evidence as to whether we can learn and store multiple context-specific priors. Seydell et al. (2010) found that observers could learn to use context-specific priors for aspect ratio where the context was specified by shape but not when context was specified by colour. They suggest that this is due to a constraint on learning contextual cues that are not plausibly related to the estimated

property (i.e., the colour of an object has a less plausible relationship to its aspect ratio than shape). Chapter 5 considered whether there is a fundamental limitation on learning colour as a contextual cue by asking whether observers could learn and use two separate light priors, each contingent on colour.

The following section provides an overview of the main findings from each of the experimental chapters presented in this thesis and considers the implications and ideas for future research.

6.2 Key Findings, Implications and Future Research

6.2.1 Haptic cues are combined with visual cues to affect perceived gloss

Chapter 2 showed that our perceptual system combines information across modalities to optimise estimates not only of geometric properties such as slant (Ernst et al., 2000), size (Ernst & Banks, 2002) and shape (Helbig & Ernst, 2007b; Wijntjes et al., 2009), but also of material properties such as gloss. Although gloss is a visual property, the perceptual system uses the haptic properties of friction and compliance to affect the visual estimate. This suggests that observers have an expectation about the glossiness of an object based on how it feels: when the object felt smooth and hard, like glass, observers were more likely to report that the object was glossy than when there were no haptic cues. This was indexed by how great a deviation in the alignment between specular highlights and shading gradients observers tolerated before making matte responses. Conversely, when the object felt rougher and softer, like rubber, observers were less tolerant of deviations in the alignment between specular highlights and shading gradients than when there were no haptic cues available.

The strong effect of haptic information on the perception of gloss found in these experiments is especially striking considering the relative paucity of haptic information available to observers. In everyday interactions, as people make judgements about objects, they typically have access to both kinaesthetic and cutaneous aspects of haptic information. Cutaneous inputs are from the mechanoreceptors and thermoreceptors in the skin and kinaesthetic inputs are from mechanoreceptors embedded in muscles, tendons and joints (Lederman & Klatzky, 2009). Using the PHANToM to present virtual haptic stimuli it is possible to present only the kinaesthetic forces which act upon a person as she explores an object and not the cutaneous elements such as fine grained texture or thermal properties. It seems likely that the effect of haptics on the visual percept of gloss noted here would be even stronger if cutaneous cues were also present in the scene, although this is a question for further investigation. Another difference between the method used here and real world

interactions was that as participants touched objects and assessed their compliance they were not able to see comparable deformations on the visual objects. This would have been more important for the softer, rubbery objects than the hard, glassy objects since glass does not visibly deform. However, for the rubbery objects, the lack of visible deformations could have reduced the sense that the felt object was one and the same as the seen object, thus reducing observers' dependency on the haptic information and leading to greater attendance to the visual information. The outcome of this would have been that there was no discernible difference between the visual and visual-haptic conditions: despite the lack of coherency in deformation, it appears that people were perceiving the felt and seen objects as one since the haptic cues did have an effect on their visual percept for both the rubber and glass conditions.

There has been little other work regarding the interaction between visual and haptic cues to material properties. Observers are able to combine information from touch and vision to improve acuity when making judgements of the roughness of surfaces made of varying grades of sandpaper (Lederman & Abbott, 1981; Heller, 1982). In these tasks the smaller the difference in grade that can be detected, the greater the acuity. These are perhaps unusual stimuli that do not reflect many normal objects; nonetheless they demonstrate that we do combine information from different senses when making judgements of material properties. In common with the aforementioned studies on visual-haptic cue combination for geometric properties, these studies of texture perception were founded on the premise that each property is accessible both visually and haptically. In most studies of cross-modal integration researchers have measured the response from each modality and then compared this with the response from both modalities together (e.g., Alais & Burr, 2004; Ernst et al., 2000; Ernst & Banks, 2002; Lederman & Abbott, 1981). To assess whether there is statistically optimal (or near optimal) integration the usual method is to compare the variance in estimates from each modality with those from the two modalities together. If there is optimal cue combination then the variance in the estimates from both cues combined will be less than, or equal to, the variance in the estimates from the least variable of the two cues. Unfortunately this method is not appropriate for assessing the cue combination in the current study of haptics and gloss, as it is not clear what it would mean for someone to explore an object haptically and estimate whether it is shiny or not. Gloss, unlike size, slant or surface roughness, is primarily a visual property, although given that haptics affects gloss perception it appears that there are haptically accessible correlates of gloss. It might be that there is a process akin to cue promotion (see Section 1.3.1.3) required to change the haptically accessible correlates into an estimate of gloss. Having unknown haptic correlates makes further quantitative analysis of these results difficult since it is not clear whether there are two estimates of gloss, one visual and one haptic that are later combined, whether there are strong fusion type interactions between the two properties or whether there is a bias (perhaps using a coupling prior) towards

interpreting the visual scene in a particular way when certain haptic information is present.

An alternative interpretation of the results presented in Chapter 2 is that the haptic stimuli introduced a response bias, at a higher cognitive level, rather than affecting the actual percept of gloss. This hypothesis could be tested in a future study by measuring discrimination thresholds for both properties individually (haptic and gloss), and using these to create stimuli that would be metameric under full optimal integration (see Section 1.4.1). The individual cues in metameric stimuli would be incongruent, that is, the haptic and visual cues would be varied in opposite directions (e.g., hard and matte or soft and glossy). Congruent stimuli would also need to be tested, in which the two cues would be varied in the same direction (e.g., hard and glossy or soft and matte). The difference in discrimination performance between congruent and incongruent stimuli, when compared with a standard stimulus, would provide a measure of the degree to which the haptic and visual cues to material are integrated at a perceptual level.

Other studies have demonstrated the complementary effect, that visual cues can affect estimates of haptic properties, using the 'material-weight illusion' (e.g., Buckingham, Cant & Goodale, 2009). Objects that had identical size and mass were rated as having different weights depending on the material they appeared to be made from: polystyrene was rated as heavier than wood which was rated as heavier than metal. Conversely, objects that had different mass (although still identical size) were perceived as having the same weight when the visual properties were manipulated appropriately: the lightest object appeared to be made of polystyrene, the middle object appeared to be made of wood and the heaviest object appeared to be made of metal. Learned associations between the visual cues and haptic material properties would seem to drive the material-weight illusion; similar learned associations between haptic and visual material properties would explain the effect presented here as well.

In addition to the novel finding that haptic cues affect perceived gloss, the results confirm previous findings that the further a specular highlight is offset from the shading gradient, the less likely it is that the object will be interpreted as shiny (Anderson & Kim, 2009; Kim et al., 2011; Beck & Prazdny, 1981). This is further evidence against the suggestion by Motoyoshi et al. (2007) that glossiness is dependent on the skewness of the luminance distribution of an image; the spatial relationships between shading and specular highlights are also important.

An alternative, more indirect way, to investigate observers' interpretation of a potential highlight would be to use highlight colour. The colour of a specular highlight usually indicates the colour of the illuminant, whereas the colour of the diffuse reflectance component depends on both the illuminant colour and the surface colour. If a bright patch in the image is interpreted as a specular highlight (i.e., the surface is

seen as glossy), the colour of the highlight can be used to discount the illuminant colour from the surface and so will change the perceived surface colour (Snyder, Doerschner & Maloney, 2005). If, on the other hand, the highlight is interpreted as a pigment change on a matte object, the colour of the diffuse shading will be presumed to be a reflection of white light from a surface of that colour. The colour of diffuse and specular reflections from the object could be manipulated, with observers asked to match the surface colour to sample colours. The advantage of having an indirect measure of gloss, like this, is that observers sometimes found it difficult to classify objects as shiny or matte whereas a colour matching task may be easier.

6.2.2 Highlight disparity cues affect perceived gloss without reversing shape percepts

The second experiment in Chapter 2 further investigated how highlight disparity affects perceived gloss. Three different highlight disparities were used: correct for the convex surface; zero relative disparity, that is, the same as the surface of the object (as though there was a pigment change); and reversed (correct for a concave object of the same size and curvature). Previous studies by Wendt et al. (2008, 2010) found that the presence of disparity affected the authenticity and strength of perceived glossiness but they did not consider negative disparities, i.e., highlight disparities consistent with a concave object.

Chapter 2 showed that there was a significant effect of highlight disparity on the percept of gloss. Objects which had a highlight disparity consistent with the other cues to convex shape looked shinier than objects where the highlight was consistent with a concave object, which in turn looked shinier than objects where the highlight was consistent with a pigment change. These results fit well with previous findings that the strength of gloss percepts is affected by highlight disparity (Wendt et al., 2008, 2010), and extend them by also considering the effect when the highlight disparity is inconsistent with the shape specified by other stereo cues. The significant difference between each of the 'convex', 'concave' and 'pigment' conditions suggests that the effect of highlight disparity on gloss perception is described by a non-symmetric function, whereby highlight disparity has the greatest effect when the sign is consistent with surface shape, still has some effect when it is inconsistent and has the least effect when it is consistent with a surface pigment change. Indeed an asymmetric model was found to fit the gloss response data from convex objects better than a symmetric model in Chapter 3. That gloss ratings for the 'concave' (reversed highlight disparity) condition would be higher than those for the 'pigment' (zero relative highlight disparity) condition is somewhat surprising as both are inconsistent with a highlight on the observed convex object. One possible explanation for the observed pattern of gloss percepts is that specular highlight disparity provides a cue to both gloss and

shape (e.g., Blake & Bülthoff, 1990, 1991); the conflict between highlight disparity (suggesting object concavity) and other cues to shape (surface disparity specifying object convexity and ambiguous shading consistent with either interpretation) may have increased shape uncertainty, potentially resulting in the perception of a concave object. In this case, the highlight would then appear to be in the correct position and so the object would appear glossy.

Chapter 3 considered whether shape conflict could explain the findings from the previous experiment as well as results reported by Blake & Bülthoff (1990, 1991). Blake & Bülthoff found that observers were sensitive to the geometry of highlight disparity when judging the shape of both convex and concave objects; observers also adjusted highlight disparity in a geometrically consistent fashion when asked to maximise the perceived gloss of a convex object. However, when maximising the perceived gloss of a concave object, observers adjusted highlight disparity such that it was consistent with a surface pigment change or a glossy, convex surface. Blake & Bülthoff attributed the deviations from a geometric model of highlight disparity to shape cue conflicts resulting in shape uncertainty.

The results presented in Chapter 3 showed that increasing shape cue reliability increased the effect of highlight disparity on gloss; this is unsurprising since the more certain the observer is about the surface disparity, the more able she is to determine precisely the relative highlight disparity. If the observer is more certain what the relative highlight disparity is, that disparity will have more effect on her gloss judgements. However, Chapter 3 also shows that even when shape is well-defined using reliable shape cues, gloss perception is not consistent with a geometric model of highlight disparity. This is in contrast with the conclusions of Blake & Bülthoff (1990, 1991).

There were some differences between the results of Chapter 2 and Chapter 3. Gloss judgements of convex surfaces in the former appeared to be inconsistent with a geometric model of highlight disparity whereas in the latter they were consistent. Methodological differences between the two experiments may go some way to explaining these discrepancies. In Chapter 2 observers made two-alternative forced choice responses between ‘shiny’ and ‘not shiny’ compared with continuous responses in Chapter 3. Gloss is not a binary concept; there are gradations which mean it is possible to say that although one material is shinier than another, both are glossy. There were also large individual differences between observers’ gloss responses for convex surfaces: despite good stereoacuity, a subset of observers tended to perceive surfaces with the incorrect sign of highlight as glossy. The perception of gloss for concave surfaces deviated from a geometric model of highlight disparity: although the size of highlight disparity mattered, the sign did not.

The difference in gloss judgements between concave and convex objects with incorrect highlight disparity suggests that observers have a better model of the geometry of highlights for convex surfaces. There are several possibilities as to why this might be. Firstly, observers may simply have less experience with concave surfaces than convex ones and so may not have learned completely the rules governing shape and material perception. Secondly, the geometry is simpler for convex surfaces - they always generate virtual highlights, regardless of where the light source is located. Concave surfaces may, depending on the position of the light source, generate either real or virtual highlight disparities; in addition to this there may be inter-reflections in concave surfaces which might further complicate the highlight disparity relationship. Training observers with glossy concave surfaces might improve their models of highlight geometry to cope with the extra complexities and/or lack of experience with concave surfaces.

The experiment presented in Chapter 3 varied horizontal disparity whilst keeping vertical disparity constant by manipulating the simulated inter-pupillary distance in both horizontal and vertical axes. Another interesting line of enquiry would be to maintain horizontal highlight disparity whilst manipulating vertical disparity to explore the effect on gloss and shape perception. At present it is only possible to vary separately and systematically the vertical and horizontal disparities of specular reflections generated by a simple object illuminated by a single point light source; stimuli like those used in Chapter 3 would be suitable to compare the effects of varying horizontal and vertical highlight disparities. More complex surfaces and light fields produce a wide range of horizontal and vertical disparities; the only systematic manipulation possible in this case is to vary the simulated horizontal inter-pupillary distance (e.g., Muryy et al., 2012), the effect of which can vary dramatically across the disparity field. The manipulation of horizontal inter-pupillary distance could equally be applied to the simpler stimuli described above, where it could be used to explore the interaction between horizontal and vertical highlight disparity in a more controlled fashion than is possible when using complex light fields.

6.2.3 Young children display evidence of optimal audio-visual cue integration

Chapter 2 and Chapter 3 advanced our understanding of cue integration in adulthood by extending the classic studies of cue integration for geometric properties to the topic of material perception. Although it seems clear that adults are able to integrate cues to improve the precision of cue estimates in a range of perceptual tasks, it is not clear when these capabilities develop. To investigate the age at which cue integration develops it was desirable to choose a topic in which adult cue integration has been well established using quantitative models, and in which the required experimental

equipment could be easily transported to allow data collection at a local school. In response to these constraints, two audio-visual computer based tasks were designed (Chapter 4) to test and compare audio-visual integration capabilities in children and adults.

The first experiment tested whether auditory stimuli can affect children's visual percepts using the bounce/stream illusion. The extent to which the rate of bouncing percepts increases in audio-visual trials compared with visual only trials can be used as an indication of the extent to which auditory information is integrated with visual information. The second experiment used the fission and fusion illusions to provide a stronger test of audio-visual integration capabilities in adults and children by making quantitative predictions of performance under two different strategies: optimal partial cue integration and a switching model.

Only one previous study had investigated the bounce/stream effect in childhood (Scheier et al., 2003) and they used pre-verbal infants so had to rely on looking time measures as a proxy for whether percepts were different when an auditory stimulus was added to the visual scene. They concluded that infants did experience the bounce/stream effect but there remains some debate as to how to interpret the looking-time measures used (Slater, 2003). Chapter 4 showed that children had a strong tendency to interpret all of the stimuli as 'bouncing', a tendency that reduced with age. The tendency for children to see all stimuli as bouncing is inconsistent with the conclusions drawn by Scheier et al. (2003): the results presented here suggest that the infants in their study may not have perceived the bounce/stream effect but were instead responding to differences in the timing of the auditory stimulus relative to the visual scene.

The tendency for the proportion of 'bouncing' percepts to reduce with age in the children studied might result from the development of attentional processes rather than cue combination. In adulthood the proportion of 'bouncing' percepts increased both with the addition of visual distractors and an additional discrimination task (Watanabe & Shimojo, 1998); it is possible that the auditory stimulus acted as a distractor that interrupted attention from the visual stimuli. Watanabe & Shimojo (1998) suggested that attention is necessary for the perception of continuous motion and that when distracted the motion percept will be disrupted - reducing the probability of a streaming percept. One way to test this possibility would be to repeat the bounce/stream experiment presented here but to include a condition in which there are visual distractors and compare the effect of additional visual and auditory stimuli on the proportion of bounce percepts. If the effect is a purely attentional one, then visual distractors should have the same effect as the auditory ones in increasing the proportion of bounce percepts in children. If, on the other hand, the increase in bouncing percepts is (partially) due to audio-visual integration, then the visual

distractors should have a lesser effect on the proportion of bounce percepts than the auditory stimuli.

The bounce/stream effect experiment provided evidence that auditory stimuli can affect visual percepts in 5-7 year old children but could not determine whether this was as a result of optimal cue integration. The second experiment in Chapter 4 used the fission and fusion illusions to test quantitative predictions of audio-visual cue integration. It provided evidence that children aged 5-7 years not only integrate auditory and visual cues but that they do so in the same way as adults: the response variance from both adults and children was better predicted by a coupling prior model of optimal integration than by a switching model.

Two previous studies have used the fission and fusion illusions to investigate whether children integrate audio-visual information, but with contradictory conclusions. Tremblay et al. (2007) found that there was no difference in the number of either fission or fusion illusions between 3 age groups in the range 5-19 years old. Despite not including an adult control group, they concluded that audio-visual cue integration mechanisms were already mature in the youngest age group (5-9 years old). Conversely, Innes-Brown et al. (2011) found that children (8-17 years old) experienced a larger fission effect than adults. They concluded that audio-visual integration is immature in 8-17 year olds. However, in both cases the methodology and analysis do not justify such strong conclusions. In particular, the size of fission and fusion effects do not provide a direct measure of audio-visual integration strategies as the reliability of the individual cue estimates may also differ between adults and children. In such a case, the same strategy might lead to different responses. To determine whether a common audio-visual integration strategy is used in childhood and adulthood it is necessary to determine how percepts in bimodal trials are related to percepts on unimodal trials.

The second experiment in Chapter 4 demonstrated the same differences in fission effects between adults and children as were found by Innes-Brown et al. (2011), in addition to differences in the strength of fusion effect. These differences could have led to the same conclusion, that audio-visual integration mechanisms are immature in 5-7 year olds. However, subsequent modelling of bimodal response variance using unimodal percepts as inputs allowed comparison of the strategies used by adults and children. Modelling showed that the response variance was better predicted by the coupling prior model than the switching model for both adults and children. The differences in number of fission and fusion illusions between adults and children can be explained by differences in the relative reliabilities of the auditory and visual cue estimates and differences in the strengths of their coupling priors, rather than differences in integration strategy. Neither model was entirely accurate; in particular, the coupling prior model failed to predict the small increase in response variance in conflict trials compared with congruent trials, for both adults and children. There were a number of experimental factors which might have contributed to the differences between the

empirical and predicted behaviour. Firstly, the permitted response range was fixed such that response distributions were truncated, and therefore asymmetric in the one and three event conditions. Additionally, the very small number of trials in each condition and the high reliability of the auditory stimuli resulted in several observers having a unimodal accuracy rate of 100%, and consequently a variance estimate of zero; this is almost certainly an underestimate. These weaknesses could be addressed in future studies by expanding the range of permitted responses; increasing the number of trials per condition; and reducing the reliability of the auditory stimuli, perhaps by increasing background noise levels. However, it would be necessary to ensure that such changes did not make the task too difficult for the youngest participants.

One additional enhancement which would further test the coupling prior model would be to collect auditory responses ('how many beeps?') as well as visual responses on bimodal trials. This would reduce the number of assumptions required to calculate the coupling prior variance, by giving a direct measure of the effect of visual stimuli on auditory responses. Measuring auditory and visual responses would also allow estimation of the auditory and visual marginal posterior distributions independently of one another, providing a means to test the prediction of the coupling prior model that these two distributions need not be identical under partial coupling.

Other studies have only found optimal integration with appropriate cue weighting in children older than those in the current study: they found that optimal integration developed at some point in the age range of 8 to 12 years old (Gori et al., 2008; Nardini et al., 2008, 2010, in press). As discussed in Section 4.1, Nardini et al. (in press) noted that most of these studies also had working memory demands (Gori et al., 2008; Nardini et al., 2008) and/or required comparison of multiple stimuli (Gori et al., 2008; Nardini et al., 2010). These additional task demands could obscure any small decreases in bimodal variance resulting from an optimal integration strategy. To counter these problems, Nardini et al. (in press) devised a task with neither working memory demands nor a 2AFC design (described further in Section 1.5). They found that, like adults, children in the age range 7 to 9 years old used an optimal integration strategy and appropriate cue weightings whereas younger (4-6 year old) and older (10-12 year old) children did not use appropriate cue weightings. A common weakness across all of these studies is that they equate optimal integration with full coupling. Both audio-haptic (Bresciani et al., 2006) and audio-visual (Chapter 4) integration are well modelled in adulthood by partial cue integration using a coupling prior to represent the strength of coupling. This is also the case for children in audio-visual integration (Chapter 4). It is possible that children in the study by Nardini et al. (in press) were able to estimate their own cue reliabilities and weight cues appropriately at 4-6 and 10-12 years old, but that they were using a partial cue integration strategy with a weaker degree of coupling than adults and 7-9 year old children. This may also

have been the case in other studies of the development of optimal cue integration (e.g., Gori et al., 2008; Nardini et al., 2008, 2010).

The coupling prior used to model the results in Chapter 4 could be applied more generally to the study of perceptual cue integration. It models the correlation between two perceptual cues and allows interactions between cues to different world properties. When two cues provide estimates of the same property, partial coupling allows estimates from each cue to have different values to one another. Another way to think about this is that partial coupling allows the possibility that the two cues are from different sources and hence the estimates from each cue may differ, despite being drawn from the same multidimensional posterior distribution: the peak of the marginal posterior distributions for each cue may be different to one another. For example, in the audio-visual partial integration described in Chapter 4 a single two-dimensional posterior distribution is generated from the joint likelihood and coupling prior distributions. The visual and auditory perceptual estimates may take different values since partial coupling allows the two marginal posterior distributions to be different.

The coupling prior could provide a method to model more complex cue interactions using the Bayesian framework. For example, in Chapter 2, haptic cues affected perceived gloss but it is possible that visual gloss cues might also affect haptic estimates of material properties (e.g., compliance or friction) and that the correlation between these very different cues to material properties could be instantiated using a coupling prior. The coupling prior represents learned correlations between perceptual measurements and although it has so far been used mainly to instantiate weak or modified weak fusion models (e.g., Bresciani et al., 2006), it can also instantiate stronger forms of fusion in which cues cannot meaningfully be promoted to the same perceptual units (e.g., Ernst, 2007).

6.2.4 Multiple context-specific light priors can be learned for shape-from-shading

In addition to cue integration strategies that help with precision of perception, sensory systems are also able to use their knowledge of environmental statistics, in the form of prior probability distributions, to improve the accuracy of perception. Relatively little is known about the priors used for material perception compared with those used for shape perception, which is why the light-from-above prior was chosen as a target for investigation in Chapter 4. The light-from above prior, that facilitates shape from shading, is known to exist (e.g., Kleffner & Ramachandran, 1992) and to be recalibrated in adulthood with a few hours of training (Adams et al., 2004, 2010). For light priors to be maximally useful they should not only be recalibrated but the visual system should learn cues to context such that the prior probability distribution is different in contexts where the illumination statistics (e.g., average lighting direction)

differ. Chapter 5 showed that the visual system can learn and selectively invoke context-specific light prior distributions for illumination direction, with each prior dependent on colour. Observers were unaware that they did this despite switching between the two different light priors on a trial-by-trial basis. Only part of the trained shift in illumination direction was learned during the training period: approximately a third of the trained shift, in common with other studies (e.g., Adams et al., 2004). There was also some generalised learning: observers showed a shift in both contexts despite only one context having been manipulated. In contrast with Seydell et al. (2010), who found that colour could not be learned as a contextual cue to the aspect ratio of shapes, the data presented in Chapter 5 shows that colour can be learned as a context-specific cue to illumination direction. Seydell et al. found that shape (e.g., ellipses vs. diamonds) could be learned as a cue to aspect ratio, even though colour was not. They suggest that the reason colour was not learned in their study is because there was no pre-existing relationship between object shape and colour. Using their framework one could conclude that colour was learned as a cue to illuminant direction because there is a plausible causal link between colour and illumination direction: the illuminant could be a coloured lightsource at a particular position. Several other studies have suggested that it is hard (or maybe impossible) to learn new cue relationships where there is no plausible ecologically valid relationship between the two variables (e.g., Michel & Jacobs, 2007; Jain et al., 2010); this is often dichotomised into types of cue relationships that can be learned (e.g., parameter learning (Michel & Jacobs, 2007) or intrinsic cues (Haijang et al., 2006; Jain et al., 2010)) and cue relationships that cannot be learned (e.g., structure learning (Michel & Jacobs, 2007) or extrinsic cues (Jain et al., 2010)). However, since there are no set criteria by which to determine whether a cue relationship is plausible, it is hard to predict in advance of training whether that relationship should be learned. It is also difficult to show with certainty that new cue relationships cannot be learned, as failure to learn within an experimental context may simply be due to strong prior knowledge that two cues are unrelated: if this is the case, then an impractically long training period might be required for learning to be detected.

The two colours trained as contextual cues in this study were red and green; although sometimes treated as such, colour is not a categorical variable but rather a continuum. An interesting question arises as to whether the trained observers learned two distinct 1-dimensional light prior distributions or whether they learned a 2-dimensional prior for the correlation between chromaticity and illuminant direction, which could be considered a coupling prior. To test this observers could be trained using the method described in Chapter 5 but with both trained colours offset from baseline (by different amounts, but in the same direction) then tested using three illuminant colours: red; green; and a colour perceptually midway between the two (yellow/orange). If a coupling prior between chromaticity and illuminant direction has been learned, it would be expected that the light prior for the orange/yellow illuminant would lie in between

the red and green light priors. If, on the other hand, two distinct light priors have been learned, the third condition should appear aligned with the baseline light prior.

6.3 Coupling Priors for Perception

The use of coupling priors together with Bayesian statistics more generally provide a conceptual framework in which to consider the findings of this thesis. Coupling priors can be used to describe cue combination - as they were in Chapter 4. There they were used to describe the influence of an auditory estimate on a visual percept. As noted in Section 6.2.3, the same conceptual framework could be applied to the finding that haptic material cues affect perceived gloss. If there is knowledge of the correlation between haptic cues and gloss, there should also be some effect of gloss on haptic material property estimates; this hypothesis could be tested using a coupling prior model. However, there are various experimental reasons why this could be difficult: the scales for each property would need to be perceptually linear; the joint posterior distribution would need to be estimated so as to know its covariance as well as the variances of the marginal posterior distributions in each dimension, requiring a response in both modalities for each trial. If the joint posterior and likelihood were assumed to be bivariate Gaussians this would restrict the form of the coupling prior to a linear path with Gaussian spread. However, it would not be constrained to the identity line, indeed, in general, for the case of cues to different world properties there is no meaningful identity line since units would be arbitrarily defined.

In addition to being used for cue combination, coupling priors may also be used to model cue recalibration. Previous studies of the development of optimal cue integration found that it developed relatively late (between 8 and 12 years old: Gori et al., 2008; Nardini et al., 2008, 2010), and suggested that this is possibly as a result of the need to calibrate cues (Gori et al., 2008; Nardini et al., 2010). In the standard Bayesian cue combination model, a combined cue estimate is the average of two (or more) cue estimates, each weighted by their reliability (see Equation 1.7 and Equation 1.8). This is mathematically equivalent to the case in which the coupling prior is a delta function along the identity line, i.e., 'full' coupling. In this scenario the perceptual system loses access to the separate cue estimates since the resulting marginal posterior distributions are identical for both cues. Since the two estimates are now identical, they cannot be used to calibrate one another. If, on the other hand, the degree of coupling is not complete (i.e., the variance of the coupling prior is greater than zero), it is possible for the two cue estimates to differ as the marginal posterior distributions are not identical (see Section 6.2.3). The difference in the cue estimates could then be used as an error signal to drive recalibration. Ernst & Di Luca (2011) propose a mathematical model for estimating and correcting for biases in sensory estimates (effectively calibrating the individual cues) based on the coupling prior model. In their model, the discrepancy

between the two coupled cue estimates is considered an optimal estimator of the combined biases of the two sensory signals; they combine this with a 'bias prior' probability distribution (representing the probability of each sensory estimate being biased) to obtain an estimate of the sensory bias from each cue. The form of the bias prior is not necessarily related to the relative reliabilities of the two cues, since it relates to systematic rather than random error; it is therefore possible for the less reliable cue to have the greater calibratory effect on the more reliable cue, if the perceptual system believes the less reliable (noisier) cue to be more accurate (less biased). One difficulty with this approach is that it is unclear what form the bias prior should take, or how a perceptual system would learn and subsequently calibrate this. Future research could usefully be directed at developing a method for estimating observers' bias priors experimentally, in order to establish whether they exist (i.e., are non-uniform) and to generate further testable predictions. For example, measuring the bias prior for a pair of cues A and B, then subsequently for cues B and C, would allow the relative probabilities of bias from cues A and C to be estimated, from which a prediction regarding relative recalibratory effects between the two cues could be made.

Ernst (2007) showed that a coupling prior could also model the learning of a new relationship between two previously unrelated cues (stiffness and luminance). In this case the perceptual system's representation of the covariance of stiffness and luminance increased from zero over the course of training in an environment in which the two cues were correlated. The learning of a new cue relationship in this way between two or more cues could be considered as either the construction of a new coupling prior or the modification of a uniform coupling prior. It is, in principle, impossible to tell the difference, in human perceptual systems, between these two possibilities since the same decisions would result in the absence of a prior as in the presence of a uniform prior. In theoretical terms these two possibilities are different as the modification of an existing relationship may be known as recalibration or parameter learning whereas the construction of a new coupling prior would be structure learning (Michel & Jacobs, 2007). However, in biological systems, where one does not know which cue relationships are already represented in the brain, the distinction between parameter and structure learning cannot be determined.

6.4 Conclusion

The experiments presented in this thesis have extended our understanding of how perceptual cues are combined and priors are recalibrated to improve the precision and accuracy of perception. Despite the importance of the perception of material properties for effective interaction with the world, this area has previously been somewhat neglected (Adelson, 2001), although there has been growing interest in recent years. The findings from Chapter 2 and Chapter 3 contribute to this body of

work, demonstrating that haptic cues can influence the perception of gloss and that although specular highlights can be a useful cue to gloss, the visual system does not possess a full geometric model of highlight disparity.

The development of cue combination strategies is another area of cue combination which has received relatively little attention: a few previous studies have found that adult-like cue integration strategies start to be used from about 8-12 years old (Gori et al., 2008; Nardini et al., 2008, 2010, in press). Using a coupling prior model, Chapter 4 showed that cue combination, at least for audio-visual cues, is already mature at 5-7 years old, although performance is quite different between adults and children due to differences in the relative reliabilities of cues and the strength of coupling used. This may also be the case for other combinations of cues such as visual-haptic or visual-visual cue integration. Maturity of cue integration strategies does not, however, preclude learning new cue relationships beyond childhood; Chapter 5 demonstrated that adults have the ability to learn and invoke multiple context-specific light priors, using illuminant colour as a contextual cue.

The twin benefits of cue interaction: (i) noise reduction through integration and (ii) bias reduction through recalibration have both been shown in this thesis to occur in adulthood, as well as the learning of new cue relationships. It has also been shown that children benefit from noise reduction by integrating cue estimates at an age where they were previously thought to be using redundant cues purely for the purpose of (re)calibration. The coupling prior model of partial cue integration (suggested by Ernst, 2006) is able to unify these processes into a common framework, the predictions of which could inform future research into human perceptual processes.

References

Adams, W. J. (2007). A common light-prior for visual search, shape and reflectance. *Journal of Vision*, *7*(11), 11, 1–7.

Adams, W. J., Banks, M. S., & van Ee, R. (2001). Adaptation to three-dimensional distortions in human vision. *Nature Neuroscience*, *4*(11), 1063–1064.

Adams, W. J., Graf, E., & Ernst, M. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058.

Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2010). Efficient visual re-calibration from either visual or haptic feedback: the importance of being wrong. *Journal of Neuroscience*, *30*(44), 14745–14749.

Adams, W. J. & Mamassian, P. (2004). Bayesian combination of ambiguous shape cues. *Journal of Vision*, *4*(10), 921–929.

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences* (pp. 339–351). Cambridge, MA: MIT Press.

Adelson, E. H. (2001). On seeing stuff: The perception of materials by humans and machines. *Proceedings of the SPIE*, *4299*, 1–12.

Adelson, E. H. & Pentland, A. (1996). The perception of shading and reflectance. In D. Knill & W. Richards (Eds.), *Perception as Bayesian inference* chapter 11, (pp. 409–424). Cambridge University Press.

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

Aloimonos, J. (1988). Shape from texture. *Biological Cybernetics*, *58*, 345–360.

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, *21*, 301–308.

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978–R983.

Anderson, B. L. & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, *9*(11), 10, 1–17.

Anderson, B. L., Marlow, P., & Kim, J. (2012). Disentangling 3D shape and perceived gloss [Abstract]. *Journal of Vision*, *12*(9), 947a.

Atkins, J. E., Jacobs, R. A., & Knill, D. C. (2003). Experience-dependent visual cue recalibration based on discrepancies between visual and haptic percepts. *Vision Research*, *43*(25), 2603–2613.

Backus, B., Jain, A., & Fuller, S. G. (2011). Cue recruitment for extrinsic signals after training with low-information stimuli [Abstract]. *Journal of Vision*, *11*(11), 983a.

Backus, B. T. (2011). Recruitment of new visual cues for perceptual appearance. In J. Trommershäuser, K. Körding, & M. S. Landy (Eds.), *Sensory Cue Integration* (pp. 101–119). Oxford University Press, USA.

Bayes, T. (1783). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.

Beck, J. & Prazdny, S. (1981). Highlights and the perception of glossiness. *Perception & Psychophysics*, *30*(4), 407–410.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berkeley, G. (1709). *An Essay Towards a New Theory of Vision*. Skinner Row: Published by Aaron Rhames for Jeremy Pepyat.

Blake, A. & Bülthoff, H. (1990). Does the brain know the physics of specular reflection? *Nature*, *343*(6254), 165–168.

Blake, A. & Bülthoff, H. (1991). Shape from specularities: computation and psychophysics. *Philosophical Transactions of the Royal Society B*, *331*, 237–252.

Blake, A., Bülthoff, H. H., & Sheinberg, D. (1996). Shape from texture: Ideal observers and the human psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian Inference* (pp. 287–321). Cambridge University Press.

Bouzit, S., Adams, W. J., & Graf, E. W. (2007). Combining specular and diffuse lighting to recover 3-D shape. *Perception*, *36*(ECVP Abstract Supplement).

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision*, *6*(5), 554–564.

Brewster, D. (1826). On the optical illusion of the conversion of cameos into intaglios, and of intaglios into cameos, with an account of other analogous phenomena. *Edinburgh Journal of Science*, *4*, 99–108.

Buckingham, G., Cant, J. S., & Goodale, M. A. (2009). Living in a material world: how visual cues to material properties affect the way that we lift objects an perceive their weight. *Journal of Neurophysiology*, *102*, 3111–3118.

Bülthoff, H. H. & Mallot, H. A. (1988). Integration of depth modules: stereo and shading. *Journal of the Optical Society of America A*, *5*(10), 1749–1758.

Bülthoff, H. H. & Mallot, H. A. (1990). Integration of stereo, shading and texture. In A. Blake & T. Troscianko (Eds.), *AI and the Eye* (pp. 119–146). John Wiley Sons Ltd., UK.

Burge, J., Ernst, M. O., & Banks, M. S. (2008). The statistical determinants of adaptation rate in human reaching. *Journal of Vision*, *8*(4), 20, 1–19.

Burt, P. & Julesz, B. (1980). Modifications of the classical notion of Panum's fusional area. *Perception*, *9*(6), 671–682.

Clark, J. J. & Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems*. Boston: Kluwer Academic Publishers.

Cutting, J. E. & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General*, *113*(2), 198–216.

Demattè, M. L., Sanabria, D., & Spence, C. (2006). Cross-modal associations between odors and colors. *Chemical Senses*, *31*, 531–538.

Demattè, M. L., Sanabria, D., Sugarman, R., & Spence, C. (2006). Cross-modal interactions between olfaction and touch. *Chemical Senses*, *31*, 291–300.

Desjardins, R. N. & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187–203.

Di Luca, M., Ernst, M. O., & Backus, B. T. (2010). Learning to use an invisible visual signal for perception. *Current Biology*, *20*, 1860–1863.

Di Luca, M., Knörlein, B., Ernst, M. O., & Harders, M. (2011). Effects of visual-haptic asynchronies and loading-unloading movements on compliance perception. *Brain Research Bulletin*, *85*, 245–259.

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology*, *21*, 2010–2016.

Drewing, K. & Jovanovic, B. (2010). Visuo-haptic length judgments in children and adults. In A. Kappers, J. van Erp, W. Bergmann Tiest, & F. van der Helm (Eds.), *Haptics: Generating and Perceiving Tangible Sensations* (pp. 438–444). Springer Berlin: Heidelberg.

Ernst, M., Banks, M. S., & Bülthoff, H. H. (2000). Touch can change visual slant perception. *Nature Neuroscience*, *3*(1), 69–73.

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In G. Knoblich, I. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Human Body Perception From The Inside Out* (pp. 105–131). New York, NY: Oxford University Press.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7, 1–14.

Ernst, M. O. (2008). Multisensory integration: A late bloomer. *Current Biology*, *18*(12), R519–R521.

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*(4), 162–169.

Ernst, M. O. & Di Luca, M. (2011). Multisensory perception: From integration to remapping. In J. Trommershäuser, K. Körding, & M. S. Landy (Eds.), *Sensory Cue Integration* (pp. 224–250). New York: Oxford University Press.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *412*, 2370–2393.

Fleming, R. W. & Bülthoff, H. H. (2005). Low-level image cues in the perception of translucent materials. *ACM Transactions on Applied Perception*, *2*(3), 346–382.

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, *3*(5), 347–368.

Fleming, R. W., Torralba, A., & Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision*, *4*(9), 798–820.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, *5*(11), 1013–1023.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Current Biology*, *18*, 694–698.

Gori, M., Sandini, G., Martinoli, C., & Burr, D. (2010). Poor haptic orientation discrimination in nonsighted children may reflect disruption of cross-sensory calibration. *Current Biology*, *20*, 223–225.

Gori, M., Tinelli, F., Sandini, G., Cioni, G., & Burr, D. (2012). Impaired visual size-discrimination in children with movement disorders. *Neuropsychologia*, *50*, 1838–1843.

Haijang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: change in visual appearance by means of Pavlovian conditioning. *Proceedings of the National Academy of Sciences*, *103*(2), 483–488.

Hartung, B. & Kersten, D. (2002). Distinguishing shiny from matte [Abstract]. *Journal of Vision*, *2*(7), 551a.

Hartung, B. & Kersten, D. (2003). How does the perception of shape interact with the perception of shiny material? [Abstract]. *Journal of Vision*, *3*(9), 59a.

Helbig, H. B. & Ernst, M. O. (2007a). Knowledge about a common source can promote visual-haptic integration. *Perception*, *36*, 1523–1533.

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

Heller, M. A. (1982). Visual and tactual texture perception: Intersensory cooperation. *Perception & Psychophysics*, *31*(4), 339–344.

Hershberger, W. (1970). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of Comparative & Physiological Psychology*, *73*, 407–411.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science*, *298*, 1627 – 1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *Journal of Vision*, *4*(12), 967–992.

Honey, R. C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Quarterly Journal of Experimental Psychology Section B*, *47*(1), 83–103.

Innes-Brown, H., Barutchu, A., Shivdasani, M. N., Crewther, D. P., Grayden, D. B., & Paolini, A. G. (2011). Susceptibility to the flash-beep illusion is increased in children compared to adults. *Developmental Science*, *14*(5), 1089–1099.

Jain, A., Fuller, S., & Backus, B. T. (2010). Absence of cue-recruitment for extrinsic signals: sounds, spots, and swirling dots fail to influence perceived 3D rotation direction after training. *PLoS One*, *5*(10), e13295.

Jimenez, J., Sundstedt, V., & Gutierrez, D. (2009). Screen-space perceptual rendering of human skin. *ACM Transactions on Applied Perception*, *6*(4), 23.

Joh, A. S., Adolph, K. E., & Campbell, M. R. (2006). Why walkers slip: shine is not a reliable cue for slippery ground. *Perception & Psychophysics*, *68*(3), 339–352.

Julesz, B. (1960). Binocular depth perception of computer-generated patterns. *The Bell System Technical Journal*, *39*(5), 1125–1162.

Julesz, B. (1964). Binocular depth perception without familiarity cues. *Science*, *145*(3630), 356–362.

Kerrigan, I. S., Adams, W., Graf, E., & Chang, A. S. (2011). Highlight disparity, surface curvature and perceived gloss [Abstract]. *Journal of Vision*, *11*(12), 373a.

Kim, J., Marlow, P., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision*, *11*(9), 4, 1–19.

Kleffner, D. A. & Ramachandran, V. (1992). On the perception of shape from shading. *Perception & Psychophysics*, *52*(1), 18–36.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(ECVP Abstract Supplement).

Knill, D. (2007). Learning Bayesian priors for depth perception. *Journal of Vision*, *7*(8), 13, 1–20.

Knill, D. C. (1998). Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. *Vision Research*, *38*, 1655–1682.

Knill, D. C., Kersten, D., & Mamassian, P. (1996a). Implications of a Bayesian formulation of visual information processing for psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian Inference* (pp. 239–286). Cambridge University Press.

Knill, D. C., Kersten, D., & Yuille, A. (1996b). Introduction: A Bayesian formulation of visual perception. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian Inference* (pp. 1–21). Cambridge University Press.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Knill, D. C. & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture for judgments of surface slant? *Vision Research*, *43*, 2539–2558.

Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Krim, J. (2002). Surface science and the atomic-scale origins of friction: what once was old is new again. *Surface Science*, *500*, 741–758.

Landy, M. S. & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A*, *18*(9), 2307–2320.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Langer, M. S. & Bülthoff, H. H. (2001). A prior for global convexity in local shape from shading. *Perception*, *30*, 403–410.

Lederman, S. J. & Abbott, S. G. (1981). Texture perception: Studies of intersensory organization using a discrepancy paradigm, and visual versus tactual psychophysics. *Journal of Experimental Psychology: Human Perception & Performance*, *7*(4), 902–915.

Lederman, S. J. & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception & Psychophysics*, *71*, 1439–1459.

Lesch, M., Chang, W.-R., & Chang, C.-C. (2008). Visually based perceptions of slipperiness: Underlying cues, consistency and relationship to coefficient of friction. *Ergonomics*, *51*(12), 1973–1983.

Longuet-Higgins, M. S. (1960). Reflection and refraction at a random moving surface (I) Patterns and paths of specular points. *Journal of the Optical Society of America*, *50*(9), 838–844.

MacLachlan, C. & Howland, H. C. (2002). Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years. *Ophthalmic and Physiological Optics*, *22*(3), 175–182.

Maia, R., D'Alba, L., & Shawkey, M. D. (2011). What makes a feather shine? A nanostructural basis for glossy black colours in feathers. *Proceedings of the Royal Society B*, *278*(1714), 1973–1980.

Mallot, H. A. (2000). *Computational Vision: Information Processing in Perception and Visual Behavior*. MIT Press.

Maloney, L. T. (2002a). Illuminant estimation as cue combination. *Journal of Vision*, *2*(6), 493–504.

Maloney, L. T. (2002b). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the Physical World: Psychological and Philosophical Issues in Percepion* (pp. 145–189). Wiley, New York.

Mamassian, P. & Goutcher, R. (2001). Prior knowledge on the illumination position. *Cognition*, *81*, B1–B9.

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic Models of the Brain: Perception and Neural Function* (pp. 13–36). Cambridge, MA: MIT Press.

Marlow, P., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision*, *11*(9), 16, 1–12.

Marr, D. & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*, 283–287.

Marroquin, J., Mitter, S., & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, *82*(397), 76–89.

Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996). Throwing while looking through prisms II. specificity and storage of multiple gaze-throw calibrations. *Brain*, *119*, 1199–1211.

McCormick, D. & Mamassian, P. (2008). What does the illusory flash look like? *Vision Research*, *48*, 63–69.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Michel, M. M. & Jacobs, R. A. (2007). Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. *Journal of Vision*, *7*(1), 4, 1–18.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247–279.

Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 237–242.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, *17*(1), 154–163.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*, 206–209.

Mulligan, J. B. (1993). Nonlinear combination rules and the perception of visual motion transparency. *Vision Research*, *33*(14), 2021–2030.

Muryy, A. A., Fleming, R. W., & Welchman, A. E. (2012). Binocular cues for glossiness [Abstract]. *Journal of Vision*, *12*(9), 869a.

Nakayama, K. & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, *257*, 1357–1363.

Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences*, *107*(39), 17041–17046.

Nardini, M., Begus, K., & Mareschal, D. (in press). Multisensory uncertainty reduction for hand localization in children and adults. *Journal of Experimental Psychology: Human Perception and Performance*.

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693.

Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science*, *9*(5), 454–464.

Nishida, S., Motoyoshi, I., Nakano, L., Li, Y., Sharan, L., & Adelson, E. (2008). Do colored highlights look like highlights? [Abstract]. *Journal of Vision*, *8*(6), 339a.

Oren, M. & Nayar, S. K. (1996). A theory of specular surface geometry. *International Journal of Computer Vision*, *24*(2), 105–124.

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *ACM Transactions on Graphics*, 55–64.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317.

Pick, H., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgements of spatial direction. *Perception & Psychophysics*, *6*, 203–205.

Poggio, T. & Torre, V. (1984). Ill-posed problems and regularization analysis in early vision. In *Proceedings of AARPA Image Understanding Workshop*, (pp. 257–263).

Pollard, S. B., Mayhew, J. E. W., & Frisby, J. P. (1985). PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, *14*(4), 449–470.

Potetz, B. & Lee, T. S. (2003). Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America A*, *20*, 1292–1303.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*, 381–410.

Ramachandran, V. (1988). Perception of shape from shading. *Nature*, *331*, 163–166.

Rock, I. & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science*, *143*, 594–596.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347–357.

Sakano, Y. & Ando, H. (2008). Effects of head motion and stereo viewing on perceived glossiness. *Journal of Vision*, *10*(9), 1–14.

Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2003). Sound induces perceptual reorganization of an ambiguous motion display in human infants. *Developmental Science*, *6*(3), 233–241.

Scholl, B. J. (2006). Innateness and (Bayesian) visual perception. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Structure and Contents* (pp. 34–53). Oxford University Press.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature*, *385*, 308.

Seydell, A., Knill, D. C., & Trommershäuser, J. (2010). Adapting internal statistical models for interpreting visual cues to depth. *Journal of Vision*, *10*(4), 1, 1–27.

Shams, L., Iwaki, S., Chawla, A., & Bhattacharya, J. (2005a). Early modulation of visual cortex by sound: an MEG study. *Neuroscience Letters*, *378*, 76–81.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*, 788.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *NeuroReport*, *12*(17), 3849–3852.

Shams, L., Ma, W. J., & Beierholm, U. (2005b). Sound-induced flash illusion as an optimal percept. *Neuroreport*, *16*(17), 1923–1927.

Slater, A. (2003). Bouncing or streaming? A commentary on Scheier, Lewkowicz and Shimojo. *Developmental Science*, *6*(3), 242.

Snyder, J. L., Doerschner, K., & Maloney, L. T. (2005). Illumination estimation in three-dimensional scenes with and without specular cues. *Journal of Vision*, *5*(10), 863–877.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, *73*, 971–995.

Sun, J. & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, *1*(3), 183–184.

Takahashi, C., Diedrichsen, J., & Watt, S. J. (2009). Integration of vision and haptics during tool use. *Journal of Vision*, *9*(6), 1–13.

Todd, J., Norman, J., & Mingolla, E. (2004). Lightness constancy in the presence of specular highlights. *Psychological Science*, *15*(1), 33–39.

Torralba, A. & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE*, *2*(8), e742.

van Doorn, A. J., Koenderink, J. J., & Wagemans, J. (2011). Light fields and shape from shading. *Journal of Vision*, *11*(3), 21, 1–21.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schrillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*, 252–258.

Watanabe, K. & Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception*, *27*, 1041–1054.

Watkins, S., Shams, L., Josephs, O., & Rees, G. (2007). Activity in human V1 follows multisensory perception. *NeuroImage*, *37*, 572–578.

Watkins, S., Shams, L., Tanaka, S., Haynes, J.-D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage*, *31*, 1247–1256.

Weiss, Y., Simoncelli, E., & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598–604.

Wendt, G., Faul, F., Ekroll, V., & Mausfeld, R. (2010). Disparity, motion and color information improve gloss constancy performance. *Journal of Vision*, *10*(9), 7, 1–17.

Wendt, G., Faul, F., & Mausfeld, R. (2008). Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision*, *8*(1), 14, 1–10.

Wheatstone, C. (1838). Contributions to the physiology of vision. Part the First. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, *128*, 371–394.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. fitting, sampling and goodness of fit. *Perception & Psychophysics*, *63*, 1293–1313.

Wijntjes, M. W. A., Volcic, R., Pont, S. C., Koenderink, J. J., & Kappers, A. M. L. (2009). Haptic perception disambiguates visual perception of 3D shape. *Experimental Brain Research*, *193*, 639–644.

Wilson, S. J. & Hutley, M. C. (1982). The optical properties of 'moth-eye' antireflection surfaces. *Journal of Modern Optics*, *29*(7), 993–1009.

Woods, A. T., Poliakoff, E., Lloyd, D. M., Kuenzel, J., Hodson, R., Gonda, H., Batchelor, J., Dijksterhuis, G. B., & Thomas, A. (2011). Effect of background noise on food perception. *Food Quality and Preference*, *22*, 42–47.

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, *33*, 2685–2696.

Yuille, A. L. & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. Knill & W. Richards (Eds.), *Perception as Bayesian Inference* (pp. 123–161). Cambridge University Press, Cambridge.