Multisensory contributions to spatial perception

Betty J. Mohler[1], Massimiliano Di Luca[1] & Heinrich H. Bülthoff[1,2]

[1] Max Planck Institute for Biological Cybernetics

[2] Korea University

Correspondence about this article should be addressed to Betty J. Mohler at MPI for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany betty.mohler@tuebingen.mpg.de, or Massimiliano Di Luca at the School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT UK, max@tuebingen.mpg.de, or Heinrich H. Bülthoff at Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea, heinrich.buelthoff@tuebingen.mpg.de

Author note:

Preparation of this article by HHB was supported by WCU (World Class University) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008). The authors would like to thank Michael Barnett-Cowan and Lewis Chuang for interesting discussions and comments on the manuscript and to Ivelina Alexandrova for creating the figures.

Abstract

This chapter discusses the sensory systems that have been shown to contribute to spatial perception – vestibular, body-based, audition, and vision. We then present how spatial information is typically multisensory and estimates are integrated within and between sensory systems. Specifically, we cover how integration is modeled in terms of probabilistic inference by analyzing several topics: fusion and segregation, strategies for sensory integration, the outcome of integration, the role of prior knowledge, and how integration leads to recalibration. Finally, we discuss two topics in high-level perception that have an intrinsic multisensory nature, self-orientation perception and object recognition.

How do we know where environmental objects are located with respect to our body? How are we are able to navigate, manipulate, and interact with the environment? In this chapter we describe how capturing sensory signals from the environment and performing internal computations achieve such goals. The first step, called “early” or “low-level” processing, is based on the functioning of feature detectors that respond selectively to elementary patterns of stimulation. Separate organs capture sensory signals and then process them separately in what we normally refer to as senses: smell, taste, touch, audition, and vision. In the first section of this chapter we present the sense modalities that provide sensory information for the perception of spatial properties such as distance, direction, and extent. Although it is hard to distinguish where early processing ends and high-level perception begins, the rest of the chapter focuses on the intermediate level of processing, which is implicitly assumed to be the a key component of several perceptual and computational theories (i.e., Gibson, 1979; Marr) and for the visual modality has been termed “mid-level vision” (see Nakayama, He, & Shimojo, 1995). In particular, we will discuss the ability of the perceptual system to specify the position and orientation of environmental objects relative to other objects and especially relatively to the observer’s body. We present computational theories and relevant scientific results on individual sense modalities and on the integration of sensory information within and across the sensory modalities. Finally, in the last section of this chapter we describe how the information processing approach has enabled a better understanding of the perceptual processes in relation to two specific high-level perceptual functions: self-orientation perception and object recognition.

1 Sensory Systems in Spatial Perception

In order to understand human spatial perception, one must first understand how sensory signals carry information about the different spatial properties. This brief overview of the human

senses highlights the contributions to spatial perception and should not be considered as an in-depth description of the sensory physiology or anatomy. Interested readers can consult Moller (2002) or Wolfe et al. (2008) for more in-depth descriptions of the sensory systems and Boron and Boulpaep (2005) for physiology.

To perceive the spatial layout of the environment and produce successful actions (e.g., hitting a nail with a hammer, knocking on a door, manipulating an object, or navigating through space), humans use several types of sensory information that are collected through different sense organs. Each sense is specialized to transduce and process information coming from one type of energy (kinetic for the body senses, air vibration for audition, photons for vision). Because of physical differences in stimuli and in the information that can be obtained, it is tempting and can be very useful to consider the sensory systems as independent modules (Fodor) even though perception typically comes from processing multiple sources of sensory information at the same time. It is also tempting to regard sensation as a process in which information is passively received and no action is required, yet we recognize that organisms interact with the environment and purposefully seek information. To understand the information available through each of the sense organs we will proceed with an overview of each sense relevant for spatial perception, but it should be kept in mind that the information available is not sensed passively and in isolation for each modality. The interactive and multisensory properties of the stimuli will become clear in the rest of the chapter.

1.1 Vestibular System

The vestibular system senses translational and rotational accelerations and thus allows us to perceive the direction of gravity relative to the body, self-motion, and changes in head orientation. This information is critical for navigation or for maintaining balance. Without a

normally functioning vestibular system humans have difficulty in stabilizing their posture and gait, as well in performing complex tasks. The vestibular portion of the inner ear contains two structures: the otolith organs, which are sensitive to linear acceleration and the three semicircular canals which are sensitive to rotational acceleration (Day & Fizpatrick, 2005). The three pairs of semicircular canals (anterior, posterior, and horizontal) are arranged symmetrically at the sides of the head and the work in a push-pull fashion: when one is stimulated, its corresponding partner is inhibited. For example while the right horizontal semicircular canal gets stimulated during head rotations to the right, the left horizontal semicircular canal gets stimulated by head rotations to the left (Day & Fizpatrick, 2005). This allows us to sense all directions of rotation. The otolithic organs (utricle, and saccule) are orthogonally oriented at the sides of the head. They are sensitive to linear accelerations as they detect the displacement of small particles of calcium carbonate which sit above small hair cells. It is well known that signals from the otoliths are ambiguous indicators of self-orientation and acceleration and that other sensory signals and previous experience are needed in order to resolve this ambiguity (i.e., MacNeilage, Banks, Berger, & Bülthoff, 2007).

The vestibular system clearly provides information relevant to spatial perception and to action. For example, the vestibular system compensates for retinal image slip created by rotations of the head. Additionally, with respect to distance or depth perception, the vestibular system provides information to the active viewer about relative angular information of the head when looking at different objects. The vestibular system may also indirectly provide information about position of the body over time (as a derivative of linear and angular acceleration over time). For recent review of the vestibular system see (Goldberg et al., 2011).

1.2 Body-based senses

The proprioceptive, kinesthetic and haptic sensory systems all involve somatosensory information and combined are often referred to as "body-based senses". *Proprioception* is the sense of the relative position of parts of the body with respect to each other. *Kinesthesia* is often used interchangeably with proprioception, but with a greater emphasis on motion. When in motion, we consider kinesthesia to contribute to our proprioceptive sense, by providing precise awareness of muscle movement and joint angles to coordinate our body movements when we are in motion in our environment. For example, it is proprioception and kinesthesia that enable us to touch the tip of our nose with our eyes closed. Because proprioceptive signals provide sensory information about the position of the limbs, there needs to be a mapping to the external environment. The point-to-point mapping of the body surfaces in the brain, which was first referred to as a sensory homunculus (Penfield & Rasmussen, 1950), enables stimuli to be perceived as occurring at a specific location. It is now commonly accepted that there is a stored model of the body, or body schema, which contains representations of the shape and contours of the human body, plan of the body surface, the location of body parts, the boundaries between body parts and their relation to each other (de Vignemont, 2005; Schwoebel & Coslett, 2005). Recent research has investigated the role of this internal model of the body in visual recognition of self (Costantini & Haggard, 2007; Tsakiris, Costantini, & Haggard, 2008) and the way that this internal model is updated as the body is in motion (Wolpert, Goodbody, & Husain, 1998). In addition, our body schema is what drives embodied or grounded cognition (Barsalou, 2008; Wilson, 2002) and it is increasingly considered to be a fundamental basis of offline cognition such as memory and language (A. M. Glenberg, 1997; Arthur M. Glenberg, 2010).

Haptics is the perception of objects through active tactile interaction. It requires two afferent subsystems, cutaneous and kinesthetic, and is generally used to refer to active manual exploration of the environment and manipulation of objects (Lederman & Klatzky, 2009b). Passive touch (cutaneous) alone is often used to refer to the sensory experience (or system in some cases), which involves passively experiencing contact on one's skin. The word haptics, on the other hand, refers to the ability to manipulate and experience the environment through active exploration. Therefore haptics naturally also requires information from kinesthetic sources, because joints and muscles also move when actively touching an object. Many scientists have measured the sensitivity of humans to distinguish a passive touch on their body which depends on factors such as age, body location, and visual experience (see Lederman & Klatzky, 2009a for a review). For example, humans demonstrate higher resolution of localization of a touch on their hand as compared to their forearm (see Lederman & Klatzky, 2009a: Figure 2). It is known that the spatial resolution of the skin is not as fine as that of the visual system, but it is better than the resolution of the auditory system (see Sherrick & Cholewiak, 1986).

1.3 Audition

The collection of vibratory energies that leads to audition begins with the outer ear, which protrudes away from the head and is shaped like a cup to direct sounds toward the tympanic membrane. This structure transmits vibrations to the inner ear, which senses vibration through specialized hair cells (Boron & Boulpaep, 2005).

Spatial information can be recovered from auditory information. The brain can compare the signals transduced by hair cells from the two ears to determine the interaural time difference (ITD) and the interaural intensity difference (IID). Sounds produced to the right of the head's midline arrive at the right ear slightly earlier and stronger. The ITD and IID together provide

localization information across the entire audible frequency range, where ITD is better at lower frequencies and IID is better for higher frequencies (see Blauert, 1997 for a review). However the information does not uniquely specify the location of a sound in 3D because sources placed along a cone around the interaural axis (the “cone of confusion”) have only very small variations of ITD and ILD. Using only ITD and ILD leads to localization errors including elevation, front-back direction, and distance of the sound source. These types of confusion can be disambiguated by monoaural information, by the environmental effects (i.e., echoes), or by moving one’s head. Monoaural information about location is created by modifying the original sounds through interactions with different anatomical parts of the head. If a sound has a wide spectrum, reflections from the outer structure of the ear (pinna), skull, and hair create characteristic modulations in magnitude and delay that can be used to locate the sound source and disambiguate ITD and ILD (see Blauert, 1997; Fay & Popper, 2005 for a review).

Information about distance is relatively scarce as compared to directional localization of sounds. Sound intensity and spectral continent can be informative about absolute distance if the source is known, but can also inform about changes in distance for unknown sources (Mershon, 1997). Room reflections are also used to estimate distance of a sound source if additional knowledge might be necessary for this (Zahorik, Brungart, & Bronkhorst, 2005).

Scientists have investigated how informative auditory information is for spatial perception, focusing primarily on perception of sound source direction (Wightman & Kistler, 1999) and distance (Zahorik, et al., 2005). Furthermore, scientists have demonstrated that humans have the ability to update their own perception of location in space using only auditory targets (Ashmead, DeFord, & Northington, 1995; Loomis, Lippa, Klatzky, & Golledge, 2002).

1.4 Vision

Nearly half of the human brain is devoted to processing visual information- a proportion that far exceeds that of the other sensory systems. Therefore, unsurprisingly there has been a great deal of research on visual sensory information for spatial cognition. There are many properties of visual stimuli that carry information about space. The retinal projection contains information about radial (i.e., two-dimensional) space -- the mapping on the retinal image corresponds with relative locations of objects in space. Although relative position of objects can be judged using only relative position on the retina (judgments for which we can achieve high precision, as measured through Vernier acuity), for absolute judgments of radial location we need to know the orientation of the eyes in space. Additional information from the muscles controlling the position and orientation of the body, neck, and eyes are necessary for such an absolute judgment (Klier & Angelaki, 2008).

For what concerns spatial information along the line of sight (distance or relative depth of objects), the situation is far more complex. The optical projection of the 3D environment onto the 2D sensitive surface of the eye does not preserve the depth dimension and such information must be recovered from the visual signals. Several sources of optical information are simultaneously available to recover depth and distance (see i.e., Boring, 1952; Cutting & Vishton, 1995). Such sources are not always sufficient to specify 3D properties and for this reason they were originally termed “cues” (Berkeley, 1709) from theatre documents where the letter Q for the word “quando” (=when) which in theater scripts indicated a trigger in response to information only hinted at. It is commonly believed that cues are processed in separate modules and the output of the computation is an estimate of the geometric properties of the environment (Bruno & Cutting, 1988; Bülthoff & Yuille, 1991b; Marr, 1982). There are many types of cues: some are available in a single static image (pictorial depth cues); others are defined by

systematic transformations of the projection (dynamic cues). Still others are available via muscular information such as convergence and accommodation (oculomotor cues). Finally, several cues depend on differences between the stimulation of the two eyes (binocular cues). Pictorial cues are optical patterns on one retinal image (without any information from kinesthetic or vestibular sense modalities) due to perspective effects (relative size, horizont ratio, relative height in the field of view, texture gradients, linear perspective, aerial perspective, see Gibson, 1979), contours, occlusion, optical distortions due to refraction, and illumination (such as shading, shadows, highlights, reflections). Dynamic cues are either due to the relative motion of objects or of their visible parts (which give rise to the Kinetic Depth Effect, Wallach & O'Connell, 1953) or to the motion of the observer in the environment (motion parallax, Ives, 1929). Other cues are available because of stereoscopic signals – differences in the image projected on the two eyes – in the form of horizontal and vertical disparities (see Howard, 2002).

Several of the visual patterns in the retinal projections require prior knowledge or additional sensory information in order to be able to infer geometric properties. For example with shading patterns, it is necessary to assume the reflectance properties of the object (Pentland, 1989), the local shape (Langer & Bülthoff, 2001), or the location of the light source (Mamassian & Goutcher, 2001) if they are not specified by other visual information (i.e., Erens, Kappers, & Koenderink, 1993). Using a single pattern to recover information about distance might not be always sufficient. Only by integrating several types of cues or information from other senses is the brain able to estimate spatial properties correctly.

2 Multisensory Integration of Spatial Information

The first section of this chapter provided an overview of the sensory systems that contribute to spatial perception. We have seen that each modality provides multiple sensory signals that are

informative about the spatial properties of the environment, like distances, angles, or shape. Here we will describe how the brain processes these sensory signals to create a unique and coherent perception of the world. The importance of this mechanism was captured by James (1890 Vol 2 p 268-9) who wrote: "… space-perception consists largely of two processes--reducing the various sense-feelings to a common measure and adding them together into the single all-including space of the real world."

We will adopt the view that in order to obtain a perceptual estimate of an environmental property (such as the shape, size of an object or the location of an event) the brain uses sensory signals that do not determine uniquely their environmental causes (real shape, size, or location in physical space). The *perceptual estimate* represents the "best guess" about the world property, but it is not guaranteed to be veridical. To maximize the chance of making a good guess, the brain uses all information available – including stored knowledge about the situation. A growing number of scientists agree that the brain can solve the problem of obtaining a percept by combining sensory signals and prior knowledge similarly to the ways of Bayesian decision theory (Bülthoff & Yuille, 1991a, 1991b; Kersten & Yuille, 2003; Knill & Richards, 1996; Körding, 2007; Mamassian, Landy, & Maloney, 2002). This way of framing the problem of perception has been referred to as indirect perception (Rock, 1997). In this framework the process of integration assumes names such as Sensor Fusion or either Cue, Multisensory, or Multimodal Integration depending on whether the information is integrated at the signal level or at the level of the estimate, and whether one or more sensory modalities are involved (Ernst & Bülthoff, 2004).

2.1 Fusion and integration

Sensory signals can be more or less independent from each other when they are sensed, and it is critical for the perceptual system to be able to distinguish whether sensory signals are produced by one environmental event or many. Signals are *independent* when the sensory noise that affects the signals, i.e. limited precision of the sensory organs and neural noise, has different causes and are thus unrelated. Signals are independent, for example, when they are sensed through different sense modalities.

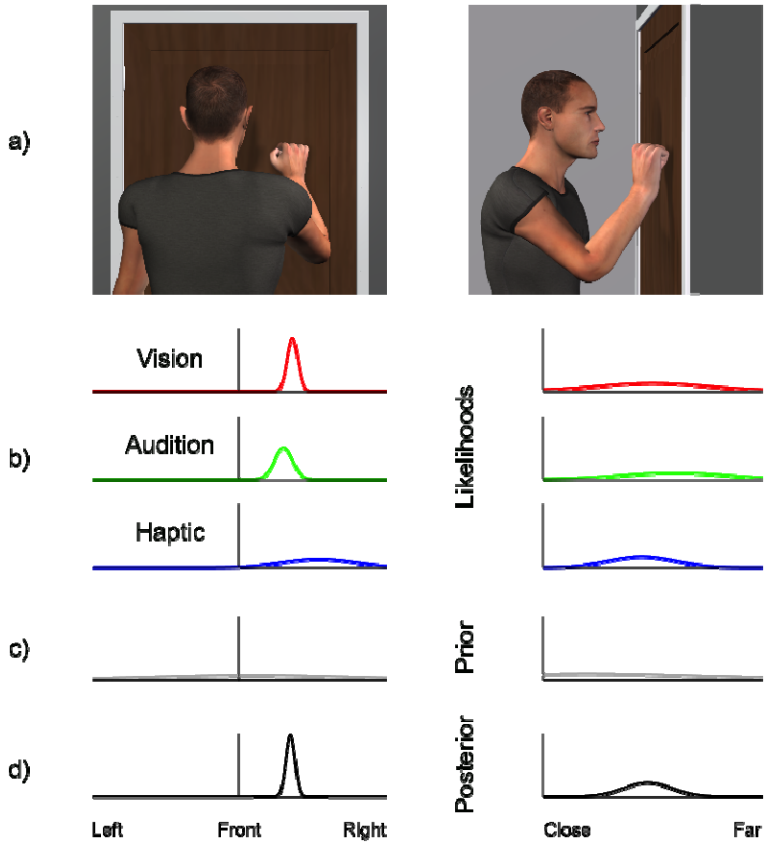

Figure 1: Visual description of the multi-sensory integration process for estimates of the location (both in terms of azimuth and distance) of one's hand knocking on a door. Visual, haptic and auditory information and prior experience with these senses all contribute to the estimate of location of one's hand when knocking on a door as is described in detail in the text.

Consider the situation in which we knock on a door with our hand (Figure 1a). The haptic, auditory, and visual signals that specify the position of where the hand hits the wood are captured by different sense organs. The neural information reaches separate brain areas, and processing is kept separate in relatively independent processing units called modules (Fodor, 1983). Nevertheless, we perceive the act of touching the wood, the sound of the knock, and the view of the hand hitting the door to be *fused* into a unified percept. If we were asked to judge the location of such a knock we could do so by estimating both the radial position along the horizontal axis and the distance (Figure 1a). Sensory signals from all three modalities could be used to perform the task because each of them is informative about location (Ernst & Bülthoff, 2004). When such types of signals are available simultaneously, they are defined to be carrying *redundant* information about location.

There appears to be some difference in how the brain deals with redundant information coming from a single sense modality and when the signals are sensed crossmodally (Hillis, Ernst, Banks, & Landy, 2002). With unimodal signals, we do not have access to the individual estimates and fusion is mandatory. For example if texture and disparity information specify conflicting information about surface slant, we are unable to estimate the two slants independently. However, if the slant is specified by visual and haptic information we can either judge the unified percept of slant, or we can judge each of the two composing slants. Note that there are many examples in which we would not want to fuse all sensory signals because they are not redundantly informative about the same environmental event. For example, if we are knocking and there is another sound in the room, it is important to keep the perception of these two events separate.

In what situations do people fuse multiple signals into one integrated percept? Multisensory integration is more likely when signals arise from approximately the same spatial location and are temporally coincident (Stein & Meredith, 1993). Radeau and Bertelson (1977, 1987) define such factors as being structural, data driven, or bottom-up although they also recognize that cognitive factors play an important role in the process. Several research lines have shown that such low-level factors are important, but they are used to make an inference about whether a common cause is responsible for the generation of all signals, a process called identity decision (Bedford, 2001; Helbig & Ernst, 2007). The extent of the integration is then a function of the probability that such a common cause exists (Körding et al., 2007; Roach, Heron, & Mcgraw, 2006; Shams & Beierholm, 2010). There are many cases in which the attribution of a common cause is purposefully used to lead to a false inference. A good example is the ventriloquism effect – the percept that speech utterances produced by the immobile lips of a puppeteer are attributed to a moving puppet. In this situation the spatial discrepancy between visual and auditory signals is disregarded leading to the percept of a single source through a process of "pairing" (Epstein, 1975; Radeau & Bertelson, 1977). The illusion is the misperception in the location of the auditory stimulus that is shifted towards the visual stimulus.

2.2 Strategies for integration

Once the brain has assessed which sensory signals belong to the same distal event and has determined that the information about location is redundant, how is this information integrated? It has been argued that during normal interaction with the environment (for example knocking on a door), one strategy could be to rely entirely on one sense to determine the perceptual outcome. For example, we could rely on vision alone and disregard audition and proprioception (termed "vetoing" by Yuille & Bülthoff, 1996). The felt position of the hand would be "captured" by

vision even when the hand is viewed through a prism that displaces it (Mon-Williams, Wann, Jenkinson, & Rushton, 1997). It might be the case that the visual modality normally leads to sufficiently precise estimates of position and hence there is no reason to incorporate other information.

It has been proposed that depending on the task to perform, there is one sensory modality that is most appropriate and the brain preferentially uses this modality (modality appropriateness hypothesis Welch, DuttonHurt, & Warren, 1986). Accordingly, in conditions with reduced illumination, the sense of hand position dominates vision (Mon-Williams, et al., 1997). However, the marked difference in precision of spatial judgments within one modality argues against this hypothesis. As we have seen in the previous section, visual judgments about angular position are much more precise than judgments in depth. Precision of proprioceptive judgments follows a different pattern, as it is affected by the geometry of the arm (van Beers, sitting, & van der Gon, 1998). This means that by using only our vision we would be very good in telling where the midline of the door is, but our estimate of the distance to the door using the same visual information would not be equally reliable. Instead, using proprioception alone we would be relatively more precise in judging distance. Thus depending on the task and context, sensory signals are differently informative and, as we discuss below, there are several advantages to considering how each signal should contribute to the final percept.

2.3 The outcome of integration

Integration fuses sensory information into a unique and coherent perception of the environment. But integration also mitigates several other types of errors that affect sensory information (see Clark & Yuille, 1990; Ernst & Banks, 2002; Ernst & Di Luca, 2011; Landy, Maloney, Johnston, & Young, 1995).

First, sensory signals are affected by noise due to the limited resolution of the sensors and by imperfect neural processing of information. When we make an estimate about the state of the world, the certainty with which we make such an estimate depends on this noise. This fact is illustrated by representing the probability (likelihood) of a state of the environment given the sensory signal across all possible states of the environment – the *likelihood distribution* for all possible locations, Figure 1b. The likelihood distributions are different for the three sensory signals because each signal is differently effective in specifying the location of the distal event. Such distributions are usually assumed to be Gaussian in shape, and their width is defined by the variance parameter. The *precision* with which it is possible to estimate a property using a signal is defined as the *reliability* of the estimate, the inverse variance of the likelihood distribution (Backus & Banks, 1999). Notice that signals in one modality might lead to estimates that differ in reliability.

Second, sensory signals might lead to estimates that are biased with respect to the true value of the property. This *bias* can either be due to random noise (and thus vary at each measurement) or it might be constant. For example, if the rotation of the head around our neck is estimated inaccurately, we can still locate objects in the environment very precisely through vision consistently across trials, but always with a constant bias. Similarly, if there is a wall on one of the side of the door we are knocking on (Figure 1a) the wall will imbalance the intensity of the sound reaching the two ears, biasing perceived location to shift towards the wall (see section I). *Accuracy* is defined as the degree to which the estimate corresponds to the true physical value of the environmental property, but unlike reliability, accuracy cannot be quantified from the current signal or the current unisensory estimate.

It has been argued that integration of multiple sources of information creates a unified percept and can increase precision and accuracy of the estimate. The Bayesian approach says that with independent sensory signals and redundant estimates, the probability of the states of the environment considering all available estimates (referred to as the *posterior distribution*) is the point-by-point multiplication of the likelihood distributions. The posterior distribution has important properties for perception (Figure 1d):

First, integration stabilizes perception in case of ambiguous estimates. For example, in some configuration of visual information (i.e., the Necker cube) or auditory cues (only ITD and ILD) sensory information is not sufficient to uniquely specify scene geometry. In this case, the likelihood distribution for the estimate of such ambiguous properties has two (or more) peaks. By integrating information within and across sensory modalities (i.e., Battaglia et al.) the point-by-point multiplication that leads to the posterior distribution can disambiguate the percept by creating a function with a single more prominent peak.

An evident difference between posterior and likelihood distributions is also their steepness. Through integration, the reliability of the posterior estimate obtained from Gaussian likelihoods increases to become the sum of the reliabilities of the individual estimates. This is the maximum improvement that can be obtained in terms of precision (when integration is statistically optimal) and for this reason it is called Maximum Likelihood Estimation (MLE). Empirical demonstrations of the increase of reliability consistent with MLE have been provided several times (see Ernst & Di Luca, 2011). One of the first studies was conducted by measuring precision of width judgments with a bar that could either be seen, touched, or seen and touched contemporarily (Ernst & Banks, 2002).

Another effect of the integration is that the peak of the posterior distribution is closer to the peak of the most reliable likelihood distribution. If all likelihood distributions are Gaussian in shape and noise is independent, the position of the peak is simply the weighted average of the position of the likelihood peaks and the weights are proportional to the reliabilities (Ernst, 2005). By recalling that reliability of visual estimates changes for lateral and depth judgments, perception should follow either the visual or the haptic estimate depending on the task and empirical results demonstrate such a close-to optimal weighting scheme (Gepshtein & Banks, 2003). Weighing of information according to reliability is also what drives the ventriloquism effect (Alais & Burr, 2004). Once an inference is made such that auditory and visual signals are generated by a common source, the perceived location of the auditory stimulus is shifted toward the visual stimulus, which is much more informative in terms of spatial location. This happens at the cost of creating the illusion of a speaking puppet.

2.4 Integration of sensory knowledge

At the beginning of this section we noted that, sensory signals are not the only source of information that can be used to make a perceptual estimate. Knowledge accumulated from previous sensory experience can influence the processing of incoming information. In the estimate of spatial properties, prior knowledge can be integrated in the posterior distribution by simply representing this knowledge as a distribution of a-priori probabilities of encountering a state of the environment, the *prior distribution*. In the knocking on the door scenario, the information represented in a prior distribution is composed of the experience of direction of knocking stimuli independent of actual sensory signals. Because we are usually the one knocking, the probability that the knock is located in front of our arm is higher than elsewhere (Figure 1c). Prior distributions are usually very shallow and deviate from normal shape, so they

exert minor influences on estimates when signals carry reliable information. But if all other sensory information is artificially reduced (e.g., with earplugs, blindfold, anesthesia, etc.) our best guess would be driven by the prior distribution. Psychophysical and sensorimotor learning experiments indicate that the perceptual outcome is consistent with independent encoding of prior information (i.e. Beierholm, Quartz, & Shams, 2009) and that the final result conforms with the predictions from the Bayesian framework (see Cheng, Shettleworth, Huttenlocher, & Rieser, 2007; Ernst & Di Luca, 2011).

Another type of prior knowledge that can be used for perception comes from the experience of multiple signals co-occurring, and as such is called a *coupling prior* (Ernst, 2005). It has been suggested that the acquisition of such a prior is what makes new signals effective in changing perception (Di Luca, Ernst, & Backus, 2010) and promotes multisensory integration (Ernst, 2007). Accordingly, research suggests that young children who have not had sufficient experience for a reliable coupling prior do not integrate multisensory information (Gori, Del Viva, Sandini, & Burr, 2008). The most frequent experience with co-occurring signals within each modality also explains why integration is stronger within than across modalities.

Such coupling priors are also important in maintaining perceptual calibration: discrepancies between the estimates could be due to either noise or to bias, and the brain may continuously assess which one is the most likely cause (for a complete discussion see Ernst & Di Luca, 2011). If in the past, the estimates have always been correlated, the cause is most likely an effect of noise. Notice that as the discrepancy increases, it can become disadvantageous to integrate. For multimodal estimates (which are not subject to mandatory fusion, see above) breakdown of integration (Gepshtein, Burge, Ernst, & Banks) with large discrepancies is a consequence of having a coupling prior whose shape deviates from Gaussian (see i.e., Roach, et al., 2006). On

the other hand, a discrepancy may be more likely due to a miscalibration when either our knowledge of the mapping is scarce, or the mapping between the estimates is weak compared to the evidence of a bias. For example, a short experience with audiovisual spatial discrepancy that induces the ventriloquist illusion also induces recalibration (Recanzone, 1998). In such cases the brain might recalibrate one or both sensory estimates. To decide in which proportion the estimates need to be recalibrate, the brain should assess the probability of bias, which is only available through prior experience with the signals (see Di Luca, Machulla, & Ernst, 2009).

In the analysis of the information available for perception, we have made the assumption that sensory signals are passively gathered from the environment. Several researchers (i.e., Marr, 1982) have subscribed to this general approach, and it has led to a wealth of scientific findings. However, perception is also achieved from dynamic sensory information and through signals that are dependent on the movement of our body in the environment. In the example of knocking on a door, the tactile, auditory and visual signals all depend on the extension of the arm. Because there is an inherent coupling of the signals in terms of their presence, magnitude, and reliability with the way in which we interact with the environment, integration should also be dependent on the way we move. For this reason, some researchers (Gibson, 1979) have criticized the assumption that perception is passively achieved and it has been proposed that perception is better conceptualized as one of the components of the perception-action cycle (Neisser, 1976). In this view, our movements are a way of picking up relevant information about the environment or the task at hand (as it happens for gaze orienting, visual search, reading, etc.) and thus to change the way we process information. Such a way of framing the perception-action loop has been successfully applied to several cases, among which the perception of the material properties of deformable objects. For example, squeezing a soft object with the hand requires the integration

of compliance estimates from multiple fingers and the one obtained with the finger that moves more (and thus the one more likely to be reliable) is also given more weight (Di Luca, 2011). Moreover, the brain treats haptic information obtained over different phases of exploratory movements differently, for example by weighting more information obtained during squeezing motion than during object release in the final percept (Di Luca, Knörlein, Ernst, & Harders, 2011).

In sum, casting the issue of sensory information integration in terms of probabilistic inferences about the state of the stimulus that generated the sensory information allows scientists to explain several aspects of the perception of spatial properties, such as the disambiguation of estimates and the increase in precision and accuracy. Such a framework also helps us to understand the relation between processing of spatial information and previous knowledge. We examine this relationship in greater detail in the next section.

3 High-level Perception

Thusfar, we have primarily discussed *low-level* or *mid-level* perception as categorized by Nakayama et al. (1995). For *high-level* perception we not only need to be able to move and understand our self-motion relative to the surrounding world, we also typically refer to the surrounding world with respect to the object themselves and not to the sensory signals that they produce and we experience. For example, we say that we heard a man knocking at the door, not that we received auditory stimulation to our ear or that we saw a door rather than a pattern of edges. The perception of one's position and orientation in the world is often referred to as self-motion perception or self-orientation perception (this research often focuses on both body-based and visual sensory information sources). The naming of objects falls under the research discipline of object and scene perception, and more specifically object recognition (most of the

research in this field focuses on visual information, although more recent research has begun to investigate auditory and haptic sensory information sources).

Humans have an impressive capability for recognizing objects, and it is not yet understood how exactly this is realized. With the rise of computer vision and the desire for machines to visually recognize objects for use in many applications, scientists from many disciplines have investigated how humans perform this function (Wallraven & Bülthoff, 2007). Recognition has often been used to refer to several high-level abilities (usually visual) including identification, categorization and discrimination of objects. Recognition of an object in this chapter is used to refer to the successful classification of an object into a specific object class (Liter & Bülthoff, 1998).

We might first ask why the process of object recognition is so difficult to understand given that it appears to be so easy for humans to perform. When performing visual object recognition, one determines whether the object he or she currently sees corresponds to an object that they have seen in the past. One possibility (albeit a brute force approach) is that we somehow store all visual stimuli associated with an object and use these stored memories to recall that object in the future. This is likely an unrealistic model because even with the enormous memory capacity that it would require, one would still be unable to experience all possible visual images generated by an object. For example, objects vary in their distance and orientation to an observer. Additionally, lighting, the context of the object (sometimes an object is alone, versus surrounded by other objects) and finally the shape of objects can vary over time. Yet humans are able to recognize newly seen objects as objects that were previously seen despite these and other variations in the scene and object. Additionally, humans constantly vary their own physical orientation with respect to the world and yet still perceive the world as upright, and are able to

recognize objects despite this change in physical viewing orientation. Therefore, before describing theories of object recognition in greater detail, we first discuss how humans determine their own orientation. These two areas of research are of course only two of the many high-level processes which have been examined from a multi-sensory integration perspective and we choose them here because of the wide interest and attention given to these topics.

3.1 Perception of Self-Orientation

When an observer moves, the sensory systems capture multiple signals: the retinal projections of the environment change, the vestibular organs sense acceleration, environmental sounds move with respect to the body, and so forth. As discussed earlier, because of the limitations of each of the sensory systems, information from multiple sense modalities is often necessary in order to navigate successfully. Vision, touch, and audition, can provide contextual information to vestibular signals for a more robust and stable representation of perceived head orientation and movement. To illustrate this, try standing on one foot with your eyes open and then try again with your eyes closed you will notice how much harder it becomes. If you then increase the amount of sensory information about changes in body orientation by lightly touching a surface with a finger or playing a localized sound, you should also notice that balance can be better maintained even when your eyes are closed (Jeka, 1997). Although information about head location and orientation comes from multiple modalities, such information may be somewhat incoherent or ambiguous. To study how multisensory integration forms a coherent perception of self-orientation, researchers introduce inconsistencies across the sensory modalities.

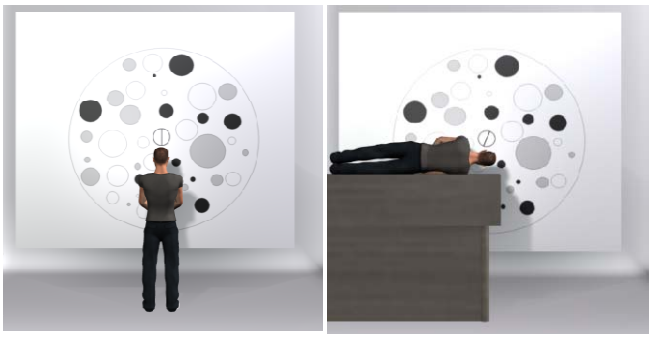

Figure 2: (Left) subjective vertical of the line in the center while standing (right) subjective vertical while a person is lying on their side is no longer a vertical line, but rather slanted in the direction of the body rotation.

The earliest investigations about the influence of head orientation on perception showed that perceived subjective vertical of viewed objects was affected by body orientation, this is the so-called “Aubert Illusion”, see Figure 2 (Asch & Witkin, 1948a; Aubert, 1861). However, with more contextual visual information it is the perceived orientation of self that is affected and not the object orientation (Asch & Witkin, 1948a, 1948b). Mittelstadt (1986) proposed that subjective vertical is obtained through a vector sum of the visual and bodily estimates based on prior experience (Dyde, Jenkin, & Harris, 2006; Mittelstaedt, 1986). It was not until much later that this was shown to be equivalent to formulating the problem in terms of Bayesian MLE (De Vrijer, Medendorp, & Van Gisbergen, 2008; Laurens & Droulez, 2007; MacNeilage, et al., 2007) where tilt estimates from the otoliths signals are combined with other sensory information (e.g. retinal line orientation). Because prior knowledge about head tilt indicates that small tilts are more likely than large tilts, this knowledge does not affect the estimates with normal postures, but it leads to errors with large deviations from vertical (De Vrijer, et al., 2008).

The subjective visual vertical is thought to be distinct from another perception of “which way is up” or what is referred to as perceptual upright (Dyde, et al., 2006) and is defined as the point at which objects are most easily recognized. Interestingly, scientists have recently found that our

own perception of self-motion (specifically self-orientation) influences the way we perceive object properties. Altering a person's sense of vertical upright and then having him or her estimate the stability of objects demonstrates the influence of self orientation perspective on object properties. It was shown that subjective vertical (and not the actual orientation of the head) could be used to predict the reported stability of an object (Barnett-Cowan, Fleming, Singh, & Bülthoff, 2011). Further, it has been shown that alterations in physical (body) and visual (object) tilt changes both allocentric (gravity orientation) and egocentric (head orientation) representations of upright, but that the vestibular system influences egocentric upright estimates more and vision influences allocentric upright estimates more (Barnett-Cowan & Harris, 2008).

3.2 Theories of Object Recognition

As we pointed out above, one critical high-level perceptual function is the recognition of previously seen objects. When considering visual object perception (and not taking into consideration body-based senses) the problem could be simplified and conceived as a problem of 2D retinal information that needs to be indexed or classified. For 2D object recognition, two approaches have been suggested: an “image-based” model, and a “structural description” model. Image-based models represent objects as a collection of viewpoint-dependent local features, while a structural description encodes objects in terms of their volumetric components and spatial relations. A review of behavioral studies concludes that while “image-based” models can explain many empirical findings, there appears to be a need for additional structural description of objects in order to explain human performance in object recognition tasks (Tarr & Bülthoff, 1998).

Although there have been many theories of object perception, Ullman (1996) divides 3-dimensional object perception approaches into three categories, 1) Invariant Properties and

Feature Space, 2) Parts and Structural Descriptions and 3) the Alignment Approach. Ullman suggests that all three approaches offer some insight into how humans perceive and recognize objects and are alone insufficient to explain human object recognition and perception. We discuss each of these three approaches in turn.

First, invariant properties and feature space theories suggest that there is a method of recognizing geometric objects that is independent of the rotation, translation, and scale of the perceived object (as well as other variations such as lighting and some aspects of shape). Specifically, supporters of this approach argue that certain properties such as area, elongation, perimeter length, and shape moments can be used to recognize objects (see, for examples, Bolles Cain, 1982; and theory in Gibson, 1979). While this approach has worked in explaining the recognition of simple objects, for more complex objects the use of such simple invariants has not proven to be useful, without combining this approach with other approaches.

Parts and structural descriptions theories, as the name implies, suggest that objects are recognized not by global properties but by their parts (a widely cited theory is Recognition by Components, Biederman, 1987). This approach assumes that all objects can be broken down into a small set of “generic” components (which are shared by all objects). These generic components could be considered the basic building blocks of objects and therefore is an attractive approach because the number of generic components is limited and therefore results in an obvious cost reduction for object recognition.

While parts and structural description approaches are quite promising, they have a number of limitations. One obvious limitation is that many objects have the same parts and are yet recognized by humans as distinct objects. Additionally, the question of which generic components should be used to distinguish objects from each other is critical to the theory, and

this varies based on the group of objects that need to be recognized. Finally, the recognition of the parts of the object needs to be solved and it is not always trivial to determine which aspects of the object makes up a part (Ullman, 1996).

Finally, the alignment approach develops the idea that recognition occurs by: (a) transforming the viewed object in a way that reduces the differences between the viewed image and the corresponding stored model of the object, and (b) comparing the transformed view with the stored model. This is to say that for all objects in the world a set of known transformations are presumed possible (scaling, position, or orientation transformations). These transformations are done in order to enable a direct comparison between the viewed object and the possible object that it might be recognized as (Ullman, 1996).

Figure 3. People are slower to recognize this shape than when it is upside down because they most often see it flipped. By turning the page upside-down it is easier to recognize that the white area is the map of the continental United States of America.

One aspect of object recognition that has received a great deal of attention is the dependence upon viewing orientation. Through many psychophysical experiments, scientists have demonstrated that object recognition is viewpoint dependent in humans (Bülthoff, Edelman, & Tarr, 1995) and in monkeys (Logothetis, Pauls, Bülthoff, & Poggio, 1994). This has been

demonstrated in a number of studies that show that recognition performance decreases as the orientation is further from the trained orientation (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Rock & Di Vita, 1987; Tarr & Pinker, 1989; Tarr & Pinker, 1991). This dependence on viewing direction has been shown even in the presence of stereo, shading and motion cues (Bülthoff & Edelman, 1992). These findings go against theories that support a structural description as well as the Alignment approach because structural theories predict no dependence on viewing direction, while the alignment approach should be sufficient to generalize over a wide range of viewing orientations, if this approach is not view dependent. Additionally, it has been shown that recognition for novel views improves after training with additional object views (Poggio & Edelman, 1990; Tarr & Pinker, 1989). For example, Bülthoff and Edelman (1992) showed that participants performed better on object recognition tasks when their orientation varied along a single axis (such as yaw or heading) which supports a 2-dimensional image combination approach to 3-dimensional object recognition (see Ullman, 1998). For a thorough and interesting review on the insights and progress on visual object perception in the past 20 years (both psychophysical, neuroscientists and clinical results) see Peissig & Tarr 2007.

Many scientists have also considered object recognition to be an active exploration process (Ernst, Lange, & Newell, 2007). The ability to generalize from a previously seen view of an object to new views of the same object involves some understanding of the spatial relations between the views. It has been shown that this process is facilitated when human observers experience walking or physically moving thereby experiencing body-based sensory information (Christou & Bülthoff, 1999; Simons, Wang, & Roddenberry, 2002; Teramoto & Riecke, 2010) or object manipulations that are congruent to the changes in views (James; et al., 2002; Meijer &

van den Broek, 2010). It has been suggested that activity during view changes could facilitate the process of 'mental rotation'.

Additionally, haptic object recognition is an active area of current research. Interestingly, it has been shown that although with visual object recognition best performance is observed from the front of a canonical view, people recognize an object best haptically when they explore the object from the opposite side of the canonical view (or the back) (Newell, Ernst, Tjan, & Bülthoff, 2001). Additionally, auditory information can also be used for object categorization (Werner & Noppeney, 2010a, 2010b) and prior experience with auditory information has also been shown to influence categorization of objects (Adam & Noppeney, 2010).

4 Conclusions

In this chapter, we have described how sensory information collected through multiple sense organs enables perception of spatial properties. We have shown that humans possess several sensory systems each attuned to one type of energy, limiting the quality of the perceptual estimate that can be obtained. Research suggests that integrating the information from the different sensory systems is a way of improving our perceptual abilities and of performing actions successfully. We have described how multisensory integration is consistent with the view that the brain attempts to maximize the extraction of information according to Bayesian accounts of perception. We have discussed the relevance of multisensory information to perception for action by describing interactive situations that require spatial information processing. Finally we have analyzed how the processing of sensory signals leads to the extraction of the information leading to high-level perceptual processing that interfaces with cognitive functions.

5 Further Suggested Reading (Annotated)

- Bülthoff, I. and Bülthoff, H. H. *Image-based recognition of biological motion, scenes and objects*, 146-176. (Ed) M.A. Peterson and G. Rhodes, Oxford University Press, New York, (2003).

  This chapter is an excellent review of what is known about image-based recognition of objects as well as biological motion and scene recognition. The authors do an excellent job discussing the difficulties at all levels of scene recognition and leave the reader impressed with human's ability to recognize objects in the surrounding world.

- Ernst MO and Bülthoff HH (2004) Merging the Senses into a Robust Percept *Trends in Cognitive Sciences* 8(4) 162-169.

  This was one of the first articles to describe multi-sensory integration and is recommended for young scientists as well as those who have studied perception for some time and want to gain a greater understanding of the theory of multi-sensory integration.

- Peissig, Jessie and Tarr, Michael J., Visual Object Recognition: Do We Know More Now than We Did 20 Years Ago?. Annual Review of Psychology, Vol. 58, January 2007. Available at SSRN: http://ssrn.com/abstract=1077345

  This review helps the reader to see what scientists have learned over the past 20 years about visual object recognition. This review discusses not only psychophysical and behavioral results, but also the neural underpinnings of how humans recognize objects visually.

- Ernst, M. O., & Di Luca, M. (2011). Multisensory perception: from integration to remapping. In J. Trommershäuser, M. S. Landy & K. Körding (Eds.), *Sensory cue integration*. Oxford, UK: Oxford University Press.

This is a more advanced and in-depth description of the processes needed for multi-sensory integration as well as a review of recent research results on the topic.

6 Bibliography

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage, 52*(4), 1592-1602.

Alais, D., & Burr, D. C. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257-262.

Asch, S. E., & Witkin, H. A. (1948a). Studies in space orientation. I. Perception of the upright with displaced visual fields. *J Exp Psychol, 38*, 325-337.

Asch, S. E., & Witkin, H. A. (1948b). Studies in space orientation: II. Perception of the upright with displaced visual fields and with body tilted. *J Exp Psychol, 38*, 455-477.

Ashmead, D. H., DeFord, L. D., & Northington, A. (1995). Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 239 –256.

Aubert, H. (1861). Eine scheinbare bedeutende Drehung von Objecten bei Neigung des Kopfes nach rechts oder links. *Virchows Archiv, 20*, 381-393.

Backus, B. T., & Banks, M. S. (1999). Estimator reliability and distance scaling in stereoscopic slant perception. *Perception, 28*(2), 217-242.

Barnett-Cowan, M., Fleming, R. W., Singh, M., & Bülthoff, H. H. (2011). Perceived Object Stability Depends on Multisensory Estimates of Gravity. *PLoS ONE, 6*(4).

Barnett-Cowan, M., & Harris, L. R. (2008). Perceived self-orientation in allocentric and egocentric space: effects of visual and physical tilt on saccadic and tactile measures. *Brain Res*, 231-243.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*.

Battaglia, P. W., Di Luca, M., Ernst, M. O., Schrater, P. R., Machulla, T., & Kersten, D. (2010). Within- and cross-modal distance information disambiguates visual size perception. *PLoS Computational Biology, 6*(3), 1-10.

Bedford, F. K. (2001). Towards a general law of numerical/object identity. *Current Psychology of Cognition*.

Beierholm, U. R., Quartz, S., & Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of Vision*.

Berkeley, G. (1709). *An essay towards a new theory of vision* (4th ed.). Dublin: Jeremy Pepyat.

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

Blauert, J. (1997). *Spatial Hearing*. Cambridge: MIT Press.

Bolles Cain. (1982). XXXXX This reference is missing.

Boring, E. G. (1952). Visual perception as invariance. *Psychological Review, 59*, 141-148.

Boron, W., & Boulpaep, E. (2005). *Medical Physiology*.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General, 117*(2), 161-170.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science, 89*, 60--64.

Bülthoff, H. H., Edelman, S., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5*, 247-260.

Bülthoff, H. H., & Yuille, A. L. (1991a). Bayesian Models for Seeing Shapes and Depth. *Comments on Theoretical Biology, 2 (4)*, 283-314.

Bülthoff, H. H., & Yuille, A. L. (1991b). Shape from X: psychophysics and computation. *Computational models of visual processing*, 305-330.

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological bulletin, 133*(4), 625-637.

Christou, C., & Bülthoff, H. H. (1999). View dependence in scene recognition after active learning. *Memory & Cognition, 27*(6), 996-1007.

Clark, J. J., & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Norwell, MA: Kluwer academic publishers.

Costantini, M., & Haggard, P. (2007). The rubber hand illusion: Sensitivity and reference frame for body ownership. *Consciousness and Cognition, 16*(2), 229-240.

Cutting, J. E., & Vishton, P. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein & S. Rogers (Eds.), *Perception of Space and Motion*. New York: Academic Press.

Day, B. L., & Fizpatrick, R. C. (2005). The Vestibular System. *Current Biology, 15*(15), R583-R586.

de Vignemont, F. (2005). Body Mereology. In G. Knoblich, I. M. Thornton, M. Grosjean & M. Shiffrar (Eds.), *Human Body Perception From the Inside Out*: Oxford University Press.

De Vrijer, M., Medendorp, W. P., & Van Gisbergen, J. A. M. (2008). Shared computational mechanism for tilt compensation accounts for biased verticality percepts in motion and pattern vision. *J Neurophysiol, 99*, 915-930.

Di Luca, M. (2011). Perceived compliance in a pinch. *Vision Research, 51(8)*, 961-967.

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

Di Luca, M., Machulla, T., & Ernst, M. O. (2009). Recalibration of multisensory simultaneity: Cross-modal transfer coincides with a change in perceptual latency. *Journal of Vision, 9*(12), 1-16.

Dyde, R. T., Jenkin, M. R., & Harris, L. R. (2006). The subjective visual vertical and the perceptual upright. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale, 173*, 612-622.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of Three-Dimensional Objects. *Vision Research, 32 2*, 2385-2400.

Epstein, W. (1975). Recalibration by pairing: A process of perceptual learning. *Perception, 4*, 59-72.

Erens, R. G., Kappers, A. M., & Koenderink, J. J. (1993). Perception of local shape from shading. *Perception and Psychophysics, 54*(2), 145-156.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. A Bayesian view on multimodal cue integration *(Chapter 6), 105-131. (Eds.) Knoblich, G., M. Grosjean, I. Thornton, M. Shiffrar, Oxford University Press, New York, NY, USA (01 2005)* (pp. 45).

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. [Research Support, Non-U.S. Gov't]. *Journal of vision, 7*(5), 7 1-14.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 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, M. S. Landy & K. Körding (Eds.), *Sensory cue integration*. Oxford, UK: Oxford University Press.

Ernst, M. O., Lange, C., & Newell, F. N. (2007). Multisensory Recognition of Actively Explored Objects. *Canadian Journal of Experimental Psychology, 61*(3), 242-253.

Fay, R., & Popper, A. (2005). Introduction to Sound Source Localization. In A. N. Popper & R. R. Fay (Eds.), *Sound Source Localization* (Vol. 25, pp. 1-5): Springer New York.

Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge: MIT Press.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology, 13*(6), 483-488.

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.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.

Glenberg, A. M. (1997). What memory is for. . *Behavioral & Brain Sciences, 20*(1), 1-55.

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(4), 586-596.

Goldberg, J. M., Wilson, V. J., Cullen, K. E., Angelaki, D. E., Broussard, D. M., Buttner-Ennever, J., et al. (2011). *The Vestibular System: A Sixth Sense*: Oxford University Press, USA.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Current biology, 18*(9), 694-698.

Helbig, H. B., & Ernst, M. O. (2007). Knowledge about a common source can promote visual – haptic integration. *Perception, 36*(10), 1523-1533.

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*(5598), 1627-1630.

Howard, I. P. (2002). *Seeing in depth*. Toronto: I Porteous.

Ives, H. E. (1929). Motion pictures in relief. *Journal of the Optical Society of America, 18*, 118-122.

James, W. (1890). *The Principles of Psychology*. Cambridge: Harvard University Press, 1983.

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

Jeka, J. J. (1997). Light touch contact as a balance aid. *Physical Therapy, 77*(5), 476.

Kersten, D., & Yuille, A. L. (2003). Bayesian models of object perception. *Curr Opin Neurobiol, 13*(2), 150-158.

Klier, E. M., & Angelaki, D. E. (2008). Spatial Updating and the Maintenance of Visual Constancy. *Neuroscience, 156*(4), 801-818.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian Inference* Cambridge University Press.

Körding, K. (2007). Decision theory: what" should" the nervous system do? *Science*.

Körding, K., Beierholm, U. R., Ma, W., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One, 2*(9), e943.

Landy, M. S., Maloney, L. T., Johnston, E., & 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.

Laurens, J., & Droulez, J. (2007). Bayesian processing of vestibular information. *Biol Cybernetics, 96*, 389-404.

Lederman, S. J., & Klatzky, R. L. (2009a). Haptic Perception: A tutorial. *Attention, Perception and Psychophysics, 71*(7), 1439-1459.

Lederman, S. J., & Klatzky, R. L. (2009b). Human Haptics. In L. R. Squire (Ed.), *Encylopedia of neuroscience* (Vol. 5, pp. 11-18). San Diego: Academic Press.

Liter, J. C., & Bülthoff, H. H. (1998). An introduction to object recognition. *Zeitschrift fÃ¼r Naturforschung, 53c*, 610-621.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*, 401-414.

Loomis, J. M., Lippa, Y., Klatzky, R. L., & Golledge, R. G. (2002). Spatial updating of locations specified by 3-D sound and spatial language. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 335-345.

MacNeilage, P. B., Banks, M. S., Berger, D. R., & Bülthoff, H. H. (2007). A Bayesian model of the disambiguation of gravitoinertial force by visual cues *Experimental Brain Research, 179*(2), 263-290.

Mamassian, P., & Goutcher, R. (2001). Prior knowledge on the illumination position. *Cognition, 81*(1), B1-B9.

Mamassian, P., Landy, M. S., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. Rao & B. A. Olshausen (Eds.), *Probabilistic models of the brain: Perception and neural function. Neural information processing series* (pp. 13-36). Cambridge, MA, US: The MIT Press.

Marr, D. (1982). *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA, USA: Freeman and Co.

Meijer, F., & van den Broek, E. (2010). Representing 3D virtual objects: interaction between visuo-spatial ability and type of exploration. *Vision Research, 50*(6), 630-635.

Mershon, D. H. (1997). Phenomenal geometry and the measurement of perceived auditory distance. In R. Gilkey & T. Anderson (Eds.), *Binaural and Spatial Hearing in Real and Virtual Environments* (pp. 257-274). New York: Erlbaum.

Mittelstaedt, H. (1986). The subjective vertical as a function of visual and extraretinal cues. *Acta Psychol (Amst), 63*, 63-85.

Moller, A. R. (2002). *Sensory Systems: Anatomy and Physiology* (1st ed.): Academic Press.

Mon-Williams, M., Wann, J. P., Jenkinson, M., & Rushton, K. (1997). Synaesthesia in the normal limb. [Research Support, Non-U.S. Gov't]. *Proceedings. Biological sciences / The Royal Society, 264*(1384), 1007-1010.

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: a critical link between lower-level and higher level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science*: M.I.T. Press.

Neisser, U. (1976). *Cognition and reality*. San Francisco: W. H. Freeman.

Newell, F. N., Ernst, M. O., Tjan, B. S., & Bülthoff, H. H. (2001). Viewpoint Dependence in Visual and Haptic Object Recognition. *Psychological Science, 12*(1), 37-42.

Penfield, W., & Rasmussen, T. (1950). *The Cerebral Cortex of Man. A Clinical Study of Localization of Function*. New York: The Macmillan Comp.

Pentland, A. (1989). Shape information from shading: A theory about human perception. *Spatial Vision, 4*(2-3), 165-182.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*, 263-266.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics, 22*(2), 137-146.

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. *Psychological Research*.

Recanzone, G. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect. *pnas, 95*(3), 869-875.

Roach, N., Heron, J., & Mcgraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc. R. Soc. Lond. B, 273*(1598), 2159-2168.

Rock, I. (1997). *Indirect Perception*: The MIT Press.

Rock, I., & Di Vita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology, 19*, 280-293.

Schwoebel, J., & Coslett, H. B. (2005). Evidence for Multiple, Distinct Representations of the Human Body. *J. Cognitive Neuroscience, 17*(4), 543-553.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*(9), 425-432.

Sherrick, C. E., & Cholewiak, R. W. (1986). Cutaneous sensitivity. In L. K. K. Boff, & J. Thomas (Ed.), *Handbook of perception and human performance* (pp. 175-180). New York: Wiley.

Simons, D. J., Wang, R. F., & Roddenberry, D. (2002). Object recognition is mediated by extraretinal information. *Perception & Psychophysics, 64*, 521-530.

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

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

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientaiton dpendence in shape recognition. *Cognitive Psychology, 21*, 233-282.

Tarr, M. J., & Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: further issues. *Psychological Science, 2*, 207-209.

Teramoto, W., & Riecke, B. E. (2010). Dynamic visual information facilitates object recognition from novel viewpoints. *Journal of Vision, 10*(13).

Tsakiris, M., Costantini, M., & Haggard, P. (2008). The role of the right temporoparietal junction in maintaining a coherent sense of one's body. *Neuropsychologia, 46*(3014), 8.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition,* : MIT Press.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition, 67*, 21-44.

van Beers, R. J., sitting, a. c., & van der Gon, J. (1998). The precision of proprioceptive position sense. *Experimental Brain Research, 122*, 367-377.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45*, 205-217.

Wallraven, C., & Bülthoff, H. H. (2007). Object Recognition in Man and Machine. In N. I. R. I. B. Osaka (Ed.) (pp. 89-104): Springer.

Welch, R., DuttonHurt, L., & Warren, D. (1986). Contributions of audition and vision to temporal rate perception. *Perception and Psychophysics, 39*(4), 294-300.

Werner, S., & Noppeney, U. (2010a). Distinct Functional Contributions of Primary Sensory and Association Areas to Audiovisual Integration in Object Categorization. *Journal of Neuroscience, 30*(7), 2662-2675.

Werner, S., & Noppeney, U. (2010b). Superadditive Responses in Superior Temporal Sulcus Predict Audiovisual Benefits in Object Categorization. *Cerebral Cortex, 20*(8), 1829-1842.

Wightman, F. L., & Kistler, D. J. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America, 105*(5), 2841-2853.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625-636.

Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klatzky, R. L., et al. (2008). *Sensation & Perception* (Second ed.).

Wolpert, D. M., Goodbody, S. J., & Husain, M. (1998). Maintaining internal representations: The role of the human superior parietal lobe. *Nature Neuroscience, 1*, 529-533.

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: Cambridge University Press.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United With Acustica, 91*, 409–420.