‘Computational modeling of multisensory object perception’

Constantin Rothkopf*
Thomas Weisswange
Jochen Triesch

Frankfurt Institute for Advanced Studies (FIAS)
Goethe University Frankfurt
Frankfurt am Main, Germany

*:corresponding author

Abstract:
Computational modeling largely based on advances in artificial intelligence and machine learning has helped furthering the understanding of some of the principles and mechanisms of multisensory object perception. Furthermore, this theoretical work has led to the development of new experimental paradigms and to important new questions. The last 20 years have seen an increasing emphasis on models that explicitly compute with uncertainties, a crucial aspect of the relation between sensory signals and states of the world. Bayesian models allow for the formulation of such relationships and also of explicit optimality criteria against which human performance can be compared. They therefore allow answering the question, how close human performance comes to a specific formulation of best performance. Maybe even more importantly, Bayesian methods allow comparing quantitatively different models by how well they account for observed data. The success of such techniques in explaining perceptual phenomena has also led to a large number of new open questions, especially about how the brain is able to perform computations that are consistent with these functional models and also about the origin of the algorithms in the brain. We briefly review some key empirical evidence of crossmodal perception and proceed to give an overview of the computational principles evident form this work. The presentation of current modeling approaches to multisensory perception considers Bayesian models, models at an intermediate level, and neural models implementing multimodal computations. Finally, this chapter specifically emphasizes current open questions in theoretical models of multisensory object perception.

Keywords:
Theoretical neuroscience, computational modeling, Bayesian modeling, multisensory processing

3.1 Introduction

The brain receives a vast number of sensory signals that relate to a multitude of external and internal states. From these signals it has to somehow compute meaningful internal representations, reach useful decisions and carry out actions. Because of the inherently probabilistic nature of all sensing processes one of their fundamental properties is their associated uncertainty. Sources of uncertainty include the neural noise due to physical processes of transduction in early stages of neural encoding, noise due to physical constraints such as unavoidable aberrations of every imaging device, and uncertainties because many environmental states can give rise to the same sensory measurement as well as many different sensory measurements can be evoked by the same state of the world. All these uncertainties render the inverse computation from sensory signals to states of the world difficult, but the brain gives humans the perception of a stable and mostly unambiguous world. Vision, audition, touch, proprioception and all other senses suggest to us a world of individual objects and well defined states. How can the brain do this?

Over the last decades ample data has been assembled in order to shed light on how the human and primate brains are able to accomplish this feat. The literature on empirical investigations of cue combination, cue integration, perception as an inference process, and many related aspects is vast (see e.g. Kersten, Mamassian, Yuille, 2004; Ernst, Bülthoff, 2004; Yuille, Kersten, 2006 for reviews). Psychophysical, neurophysiological, and imaging studies have quantified human and primate performance and knowledge on the neural implementation has accumulated. Nevertheless, the question on how the brain merges sensory inputs into complete percepts offers many unsolved problems. Sensory processing has been traditionally thought to be separate in the respective modalities. The segregation has been applied even within modalities, as e.g. in the separation of ventral and dorsal streams in vision, promoting a hierarchical and modular view of sensory processing. But recent experimental results have emphasized the multimodal processing of sensory stimuli even in areas previously regarded as unimodal.

Theoretical work largely based on advances in artificial intelligence and machine learning has not only furthered the understanding of some of the principles and mechanisms of this process but also led to important questions and new experimental paradigms. The last 20 years have seen an increasing emphasis on Bayesian techniques in multimodal perception, mostly because such models explicitly represent uncertainties, a crucial aspect of the relation between sensory signals and states of the world. Bayesian models allow for the formulation of such relationships and also of explicit optimality criteria against which human performance can be compared. They therefore allow answering the question, how close human performance comes to a specific formulation of best performance. Maybe even more importantly, Bayesian methods allow comparing quantitatively different models by how well they account for observed data.

The success of Bayesian techniques in explaining a large variety of perceptual phenomena has also led to a large number of additional open questions, especially about how the brain is able to perform computations that are consistent with the functional models and also about the origin of these models. For this reason, and because comprehensive review articles of perceptual processes and their modeling have been published in the past years (see e.g. Kersten, Mamassian, Yuille, 2004; Knill, Pouget, 2004; Ernst, Bülthoff, 2004; Yuille, Kersten, 2006; Shams & Seitz,

2008), this book chapter specifically emphasizes open questions in theoretical models of multisensory object perception.

3.2 Empirical evidence for Crossmodal Object Perception

Considerable evidence has been collected over the last century supporting the fact that primates use input from multiple sensory modalities in sophisticated ways when perceiving their environment. While other chapters in this book focus explicitly on neurophysiological, psychophysical, and imaging studies demonstrating this ability, a few key studies are listed here with an emphasis on results that have guided theoretical work on modeling the processes manifest in crossmodal object perception.

Traditionally, it was assumed that sensory inputs mostly are processed separately and that separate cortical and subcortical regions are specialized in the integration of the distinct modalities and that such computations were late in the processing hierarchy (Jones, Powell, 1970; Felleman, Van Essen, 1991). It was assumed, that multisensory processing was confined to specific areas in so-called later stages of the processing hierarchy, a hypothesis that is reflected by the fact that such cortical areas were termed ‘associative cortices’.

Multimodal processes had been reported early on (von Schiller, 1932) and behavioral evidence for multisensory integration was shown all along, e.g. it was shown that it can speed up reaction times (Hershenson, 1962; Gielen et al., 1983), it can improve detection of faint stimuli (Frens, Van Opstal, 1995), and it can change percepts in a strong way as in the ventriloquist illusion (Pick et al., 1969), the McGurk effect (McGurk, Mc Donald, 1976), and the parchment skin illusion (Jousmaki, Hari, 1998). These studies were paralleled by neurophysiological investigations demonstrating multisensory processing in primates (Lomo, Mollica, 1959; Murata, Bach-y-Rita,1965, Bruce at al., 1981). While most of these studies were able do describe these multimodal effects qualitatively, during the last 20 years, psychophysics together with quantitative models of computations involving uncertainties had a prominent role in the understanding of multisensory perception. This was achieved by quantifying how behavior reflects quantities describing the stimuli in different modalities across a wide range of tasks. One of the main stories emerging from these investigations was the central role of uncertainty in explaining observed psychophysical performance.

Neurophysiological investigations of multisensory processing also shifted from an early emphasis on the segregation of sensory processing in separate processing streams to the rediscovery of multimodal integration in primary sensory areas. Evidence now has been found that responses of neurons in so-called early sensory areas can indeed be modulated by stimuli from other modalities.
Now, multisensory responses in early areas have been found repeatedly using a wide variety of analysis techniques including, blood oxygenation level dependent (BOLD) signals in functional magnetic resonance imaging (fMRI) (Calvert et al., 1997), event related potential (ERP) (Foxe et al., 2000), single cell recordings in macaque (Schroeder, Foxe, 2002; Ghazanfar et al., 2005), in anesthetized ferrets (Bizley et al., 2006), and in awake behaving monkeys (Ghazanfar et al., 2005). Such interactions are not restricted to so called early sensory areas, but have instead been shown to be common throughout the cortex including audio-visual responses in inferotemporal cortex (IT) (Poremba et al., 2003; Gibson, Maunsell, 1997), auditory

responses in cat visual cortex (Morrell, 1972), lateral occipital cortex (LOC) responses during tactile perception (James et al. 2002), mediotemporal (MT) activation from tactile motion (Hagen et al., 2002), visual and auditory responses in somatosensory areas (Zhou, Fuster, 2000). Furthermore, plasticity of multisensory processing can lead to remapping between modalities, as in congenitally blind and deaf individuals (Sadato et al., 1996; Kujala et al., 1995; Finney et al., 2001).

3.3 Computational principles evident from the experimental data

Theoretical modeling inherently needs to abstract from the full richness of the studied system. One of the central tools in developing models of multisensory object perception has been the formalization of the environment in terms of cues. Although a definition of a cue is highly difficult and ambiguous, the idea is to take a physical entity in the world as given and attribute it certain properties such as size, reflectance, weight, and surface structure. These measurable physical states of the world are unknown to the brain, which instead obtains signals from its sensory apparatus. Individual variables that can be recovered from this sensory input and which somehow reflect the physical state of the world are termed cues. Thus, contrast, spatial frequency, pressure sensation, sensation of acceleration, loudness have all been considered as cues. A multitude of cues have been considered in the modality of vision, where more than a dozen cues are known for depth alone including disparity, occlusion, texture gradients, linear perspective. Using this formalism, a variety of different tasks and modes of sensory processing have been distinguished.

3.3.1 Sensory integration

When there are two cues available for inferring a quantity of interest, it is advantageous to use both measurements in order to infer the unknown cause, because the common cause will usually influence both measurements in a structured way. The hope is, that by somehow combining the two measurements in a sensible way, the uncertainty in the unknown quantity can be reduced by taking advantage of this structure. This operation is called cue integration and is certainly the best studied multimodal inference task.

A wide variety of experiments have been devised to test how individual cues are combined to a single estimate of the unknown quantity. While some experiments have used different cues within the same modality, such as different depth cues in vision (Jacobs, 1999; Knill, Saunders, 2003), there are a variety of investigations of how such integration is carried out across different modalities, including audio-visual (Battaglia et al., 2003), visual-haptic (Ernst, Banks, 2002), and visual-proprioceptive (van Beers, 1999) tasks. Trimodal cue integration has also been studied as e.g. in Wozny et al. (2008). The results of these investigations have demonstrated that subjects combine the individual measurements in order to increase the reliability of their estimate of the unobserved quantity.

3.3.2 Sensory combination

The literature on multisensory processing has made the distinction between sensory integration and sensory combination (e.g. Ernst and Bülthoff, 2004). This distinguishes between cases in which different sensory signals are commensurate in the sense that they are somehow represented in the same coordinate frame, have the same units, and are of the same type, i.e. discrete or continuous, and cases in which they are not. Ernst and Bülthoff (2004) provide an example of the latter case of sensory combination involving visual and haptic information to complement each

other. When touching and viewing an object, visual and haptic information are often obtained from different sides of the object, where vision codes for the front of an object while haptics codes for the back of an object (Newell, Ernst, Tjan, Bülthoff, 2001). These different cues that are not commensurate can nevertheless be combined to aid in object perception.

3.3.3 Integration of measurements with prior knowledge

As mentioned previously, one of the central problems of the sensory apparatus is that the signals that it obtains are ambiguous with respect to what caused them. Because of this ambiguity it is advantageous to make use of the regularities encountered in the world, as these can be used to bias the interpretation away from very unlikely causes. Direct evidence that such biases are at work in every day perception comes from visual illusions involving shading. When viewing spherical objects whose upper part is brighter these are interpreted as emerging from the plane. This percept is very stable, although the image itself is highly ambiguous. The sphere could be interpreted as being convex and illuminated from above but also as being concave and illuminated from below. The explanation for this perceptual phenomenon is that the visual system biases its interpretation of such ambiguous input towards a scene configuration in which the light source is located above the object instead of below the object, because this is the more common configuration under natural circumstances. Furthermore, Mamassian and Landy (2001) showed how two priors can interact, in their case priors for the direction of illumination and for the viewpoint of the observer.

3.3.4 Explaining away

There are perceptual tasks involving several independent causes and different cues. In such situations, explaining away describes the fact that one of the causes for observed data becomes less probable when new data is observed, which makes an alternative cause more probable. Another way of viewing this phenomenon is that causes compete for an explanation of the data. As an example, consider a stimulus with two rectangular image regions both containing the same luminance gradient, which are placed next to each other. These two regions give rise to an illusion in which the lightness at the boundary between the two regions is perceived as being different across the same positions on the identical brightness distributions. Knill and Kersten (1991) reported that placing two ellipses at the bottom of the two rectangular regions strongly reduced this illusion as the ensemble image is now perceived as two equally shaped cylinders illuminated from the side. This additional scene parameter explains away the previous interpretation.

A more elaborate example involving visual and haptic measurements comes from a study by Battaglia, Schrater, and Kersten (2005). Subjects were intercepting moving targets based on monocular visual information in a virtual reality environment. When observing the target object along its trajectory, the uncertainty about the true object size transfers to uncertainty about the distance of the object and thereby renders interception uncertain, too. To test for the possibility of additional measurements to disambiguate the causes of the target's apparent size, subjects had to catch the target in some trials only based on the visual information or in different trials using visual and haptic information. The data showed that interception performance was indeed better if subjects had previously touched the target, suggesting that the additional haptic measurement enabled subjects to explain away the influence of ball size on image size leading to a better estimate of target distance.

While there is evidence for perceptual processes in which a cue is able to bias the percept away from an alternative percept rendered more likely by another cue, the literature also reports cases in which a disambiguating cue does not explain away an alternative cause. Mamassian, Knill, and Kersten (1998) report such cases involving cast shadows. In one such example, a folded paper card is imaged from above in such a way, that its shape can either be perceived as a W or as an M. The addition of a pencil laying on top of the folded card together with its shadow falling on the card should in principle exclude one of the possible shapes but fails to do so. Instead, the card is still ambiguous and can be seen to flip between both interpretations.

3.3.5 Using the appropriate model

Most of the classical cue combination studies were designed in such a way that the model describing the inference process in a particular experiment is known or controlled by the experimenter and the correct relationship between observed quantities and the unobserved causes are fully specified within that model. Crucially, in most of these studies, the sensory input is always interpreted using this single model. Some recent experiments have investigated cases in which observations may have different interpretations under different models.

Wallace et al. (2007) and Hairston et al. (2004) considered the case of a paired auditory and visual stimulus. Here, a briefly played tone and a light flash may be interpreted as having a common cause, if the positions of the two signals are close enough in space, but may also be perceived as originating from two different sources, if their spatial separation is sufficiently large. The perceptual system may select somehow, whether the appropriate model upon which the inference is to be based is that of a single common source for the two signals or whether it is more likely that the signals came from two separate sources, which is the approach taken by Koerding et al. (2007). Alternatively, it may compute the a posterior distribution of position according to the two models separately and then integrate these by somehow taking into account the full uncertainties of each models' appropriateness in explaining the observations.

3.3.6 Decision making

While subjects have an implicit reward structure when they are instructed to carry out a perceptual task in psychophysical experiments in laboratory settings, it is also important to consider different explicit reward structures. Trommershäuser, Maloney, Landy (2003) varied the reward structure in a series of fast pointing tasks and demonstrated that human subjects not only take their respective motor uncertainties into account, but also adjust their pointing movements so as to maximize the average reward. In their case, reward was made explicit through the amount of monetary compensation for the subject's participation based on their performance. These results show that the human brain is able to not only integrate sensory cues by taking their respective uncertainties into account, but that it can also apply different cost functions to the performed inferences when taking decisions.

3.3.7 Learning of cue reliabilities and priors

The above mentioned perceptual experiments varied the reliabilities of individual cues within the stimulus, i.e. a visual cue can be made more uncertain by reducing its contrast, or by adding noise to its spatial distribution. A different form of uncertainty can be manipulated when controlling the reliabilities across trials.

Such experiments often use virtual reality environments, as these allow for the full control of different stimulus modalities and allow controlling cue conflicts in a precise manner. Such experiments have shown that observers can learn the reliabilities of

individual cues through extensive training (Jacobs, Fine, 1999), and that cues in the haptic modality can be used as standard against which the reliabilities of two visual cues are judged (Atkins, Fiser, Jacobs, 2001). Furthermore, humans can quickly adjust to changing reliabilities of individual cues (Triesch, Ballard, Jacobs, 2002) across trials. So, if the system is able to reweigh individual cues by their reliabilities, it must somehow compute with the individual estimates over time.

3.3.8 Task dependence

While a visual scene contains a vast amount of information in terms of all possible variables that a perceptual system could infer from it, in a specific task only few variables really matter to obtain a judgment or plan an action. When guiding the reach of a hand in grasping a cup, it may not be necessary to infer the source of illumination or the slant of the surface on which the cup rests, but it may well be necessary to infer the shape and orientation of the cup's handle. But visual cues in such scenes may depend on both the light source and the slant of the surface. By contrast, for placing a cup on an inclined surface the slant of the surface is an important scene variable. Thus, dependent on the task at hand, some of the quantities determining the visual scene may be of direct interest and need to be inferred whereas in other tasks they don't provide important information. Explicit investigations of such task dependencies have been done by Schrater and Kersten, (2000) and Greenwald and Knill (2009).

3.3.9 Developmental learning

It may be important for the choice of the modeling framework to consider the question, whether multisensory object perception is learnt over developmental timescales or whether it is innate. Developmental studies have shown a wide range of phenomena in multisensory processing ranging from cue integration being present at an early age (Lewkowicz, 2000) to cases in which young children up to the age of 8 years did not integrate cues (Gori,2008; Nardini, 2006). Gori et al. (2008) have reported that children instead relied on only a single modality's cue when integrating visual and haptic information would have been advantageous. Which modality they relied on depended on the specific task. The range of these different results has suggested that at least some aspects of multisensory object perception are learned over developmental timescales.

3.4 Models of Multimodal Object Perception

Computational models of cognitive processes such as multimodal object perception can be built at different levels of description. It may be helpful to adopt the notion of Marr's (1982) three levels of description in order to distinguish current models of multimodal perception. At the level of *computational theory*, it can be asked what the goal of the perceptual system's computations is. While it is difficult in principle to separate the perceptual system from the goals of the entire living system, the currently dominating view is that the perceptual system aims at inferring the causes of the sensory input. This idea is often attributed to von Helmholtz (1867) but really can be traced back to Al Hazen around the year 1000 (Smith, 2001). With uncertainty about these causes being the predominant factor these computations are therefore considered to perform inference and the computational framework for modeling multimodal object perception becomes that of an ideal observer. This theoretical construct answers the question: how well could any computing device possibly do on the inference problem at hand? The problem is solved in an optimal way without any restrictions on how such computations could actually be implemented in the primate

brain. In a Bayesian framework this entails computing a probability distribution over the causes given the sensory signals. Finally, if a decision has to be taken in a normative way, such a distribution has to be combined with a cost function, which specifies the cost for all possible errors between the true value of a cause and its estimate.

At the *algorithmic level*, one may consider different approaches of how to solve an inferential task and how the sensory uncertainties might be represented. The area of machine learning has made significant advances in recent years (MacKay, 2003; Bishop, 2006) and has brought about a multitude of different algorithmic approaches on how to solve inference problems. Nevertheless, it is still an active research area with many open problems. An important lesson from this field is that even with fairly simple looking problems that involve only a few quantities that need to be represented probabilistically the inference problem can quickly become analytically intractable. This requires using approximations in order to obtain results. Indeed, many questions still remain regarding how uncertainties are represented in the brain, how computations involving uncertainties could be carried out, and which representations and approximations are most suitable for implementation in biological system.

Finally, at the *implementation level* the question is how the neural substrate may be able to implement multisensory perception. This remains a daunting problem, given that even the nature of the neural code is still debated and many question regarding the biophysical and neurochemical processes in the nervous system are still unanswered. More importantly, while more and more tools with high sophistication become available such as transcranial direct current stimulation, in-vivo fast-scan cyclic voltammetry, the use of transgenic species, photolysis by two-photon excitation, use of neurotropic viruses, the data these methods produce rarely clarify their function in terms of the computations the brain is implementing. Nevertheless, all the approaches to understanding how the primate brain accomplishes multisensory object perception need to take into account, what type of computations the brain must be able to accomplish based on the observed psychophysical evidence. In the following we describe the most prominent directions of such computational models.

3.4.1 Ideal observer models of multimodal object perception

The ideal observer framework formulates an inference task by specifying the involved quantities as random variables and expressing their probabilistic relationship in a generative model. Such generative models describe the process of how a particular sensory variable is caused by attributes of the world. When considering a visual scene, the illumination source and the reflectance properties of an object are causes for the appearance of the object in the image. The generative model also needs to specify the exact probabilistic relationship between the individual variables. When the generative model has been specified, the configuration of a scene determines which of the variables in the generative model are observed and which are hidden or latent. This will determine the statistical dependencies between the involved quantities and accordingly how the uncertainties between individual variables are related. The task at hand will furthermore determine which variable needs to be computed. While in some circumstances a particular variable may be the goal of the computation, it may be discounted in other cases. A convenient way of depicting the statistical relationships between individual variables is shown in figure 1, in which observed variables are depicted as shaded circles, non-observed or latent variables are shown

in a clear circle, and arrows represent causal relationships (see Pearl, 1988; Bishop, 2006).

Finally, when considering the full decision process based on the inference, it is necessary to apply a cost function that assigns a cost to errors in each variable dimension that is estimated. In the following, all these computations are described in more detail. The Bayesian framework can accommodate many of the above computational principles evident from the psychophysical literature. Cue combination, cue integration, integration with prior knowledge, explaining away, and certain forms of task dependence all fit in this framework (see Kersten, Mamassian, Yuille, 2004 for a review).

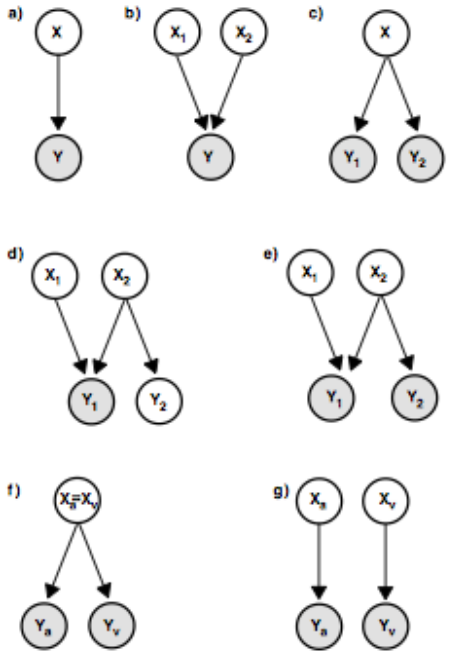

Figure 1: Representation of different statistical dependencies between non-observed (latent) scene parameters, shown as clear circles, and measurement variables, shown as shaded circles, as Bayesian nets. See text for more detail.

The essence of the Bayesian approach is to formulate a model that explicitly represents the uncertainties as probability distributions for scene variables and sensory variables that are statistically dependent on the scene variables that need to be estimated in the specific task. In the most general form, a scene variable such as the shape of an object is related probabilistically to a particular appearance image of that shape, because many different shapes can give rise to the same image or a single shape can result in many different images and because of the additional noise introduced by the imaging and transduction processes. Such a situation is depicted in figure 1a), where the latent variable X corresponds to the shape and the observed variable Y corresponds to the sensed image. Figure 1b) corresponds to the case in which two latent variables, such as shape and illumination direction are made explicit in their influence on the observed image.

As an example for multimodal object recognition we will consider the laboratory version of an audio-visual orienting task (Thomas, 1941), i.e. the task of inferring the position of an object, which is only made apparent through a brief light flash and a brief tone (Alais, Burr, 2004). In such a setting the observer has the task of inferring

the position X from two measurements, the auditory cue Y_a and the visual cue Y_v, as depicted in figure 1c). The Bayesian view is that there is not enough certainty in order to know the position of the object exactly, but that instead one needs to report the full probability distribution representing how likely it is that the object is at a particular position X=x given the two observations. Accordingly, the ideal observer obtains two measurements Y_a and Y_v and now infers the full probability distribution over all possible positions X=x so that we assign a probability to each position of the source being at that location. This probability distribution can be computed using Bayes theorem:

$$P(X|Y_a, Y_v) = \frac{P(Y_a, Y_v|X)P(X)}{P(Y_a, Y_v)} \quad (1)$$

The term on the left hand side of the equation is the posterior probability that the object is at some position X=x given the numerical values of the actually observed auditory and visual measurements. Note that this posterior distribution may have a complex structure and that there are a variety of choices of how to obtain a decision from such a distribution, if only a single percept or a single target for an action is required.

The right hand side contains first the term $P(Y_a,Y_v|X)$, which in the generative model view fully specifies how a single cause, or in the case of a vector valued X multiple causes, give rise to the observed signals. In this view, a cause takes on a particular value X=x and the term $P(Y_a,Y_v|X)$ specifies the generative process by which observable quantities are generated. This may be by linear superposition of individual causes or by some geometric dependency of visual features on the causing scene parameters. The probabilistic nature of this term may reflect the fact that multiple different causes can lead to the same scene parameters. It can also reflect the sensory noise that renders the observation stochastic and therefore the true value of the scene variable in question uncertain. In the chosen example of audio-visual localization, this term fully specifies the uncertainty in the measurements of Y_a and Y_v when the true object location is X=x, i.e. it assigns a probability to each measurement pair of Ya and Yb when the true value of X is x. Or, in the frequentist view, it describes the variability in the measurements of Y_a and Y_v on repeated trials, when the true value of X equals x.

On the other hand, when Y_a and Y_v have been observed then the expression $P(Y_a,Y_v|X)$ is viewed as a function of X and is often called the likelihood function. Note that the posterior distribution is a proper probability distribution given by conditioning on the observed values Y_a and Y_v. The likelihood function instead is conditioned on the unknown variable X. It expresses the relationship between observing a particular value for the two cues when a certain true position of the object is given. This term therefore fully describes the uncertainty associated with a particular measurement, i.e. it specifies the likelihood of sensing the actually observed values Y_a and Y_v for all possible values of X. Therefore, it may not be a proper probability distribution summing to one.

The prior probability p(X) has the special role of representing additional knowledge. In the specific case of audio visual localization this may be knowledge about common and unusual positions of the object emitting the sensory signals. While in some experimental conditions this distribution can be uniform so that it has no additional effect on the likelihood of the position of the object, the prior is often important for more naturalistic stimuli, where the unknown property of the object causing the

observations follows strong regularities encountered in the natural environment. Therefore, it may be very advantageous to use this additional knowledge in inferring the causes of the sensed stimuli, especially for large uncertainties in the sensory stimuli. In the provided example, the laboratory setup is usually such that the source of the auditory and visual stimuli are equiprobable within a predefined range, but in a natural environment it may very well be that additional priors can be taken into account. One example for such regularities is, that the sounds of birds may well have an associated prior with higher probability in the upper hemisphere of the visual field. Finally, the term $P(Y_a,Y_v)$ is not dependent on the value of the variable X but is a normalization factor that is required in order for the posterior to be a proper probability distribution summing to 1.

The above equation can be simplified under the assumption that the uncertainties in the two estimates are conditionally independent given the position of the object. One way of thinking about this, is that the noise in the auditory cue is not dependent on the noise in the visual cue. More generally, this means that if the true position of the object is known, the uncertainties in the visual and auditory cue do not influence each other. Under this independence assumption the above equation can be rewritten as:

$$P(X|Y_a, Y_v) = \frac{P(Y_a|X)P(Y_v|X)P(X)}{P(Y_a, Y_v)} \tag{2}$$

The advantage of this expression compared to equation (1) is that due to this factorization, instead of having to characterize the distribution $P(Y_a,Y_v|X)$ it is only necessary to know $P(Y_a|X)$ and $P(Y_v|X)$. The advantage lies in the fact that the joint distribution is three dimensional and would require to specify a probability value for each combination of X, Y_a, and Y_b whereas the two factored distributions $P(Y_a|X)$ and $P(Y_v|X)$ are only two dimensional. These types of independences are of central importance in Bayesian inference involving Bayes nets (Pearl, 1988; Bishop, 2006) and become particularly important when many random variables are involved and when variables change over time. Accordingly, in equation (2) the ideal Bayesian observer will base its decision on a quantity that is proportional to the product of the three terms in the numerator on the right hand side. Again, note that this quantity expresses the probability given to each of all possible values of the variable X, the position of the object. For the particular case of audio-visual object localization, the above simplification means, that only the uncertainties in object localization when either a tone alone or a flash alone are presented need to be measured in order to calculate the posterior.

Depending on the particular scene parameters and measurements considered in an inference process, it is necessary to specify the mathematical form of the involved probability distributions. A computationally particularly straightforward case is that of cue combination when the individual cues have independent additive Gaussian noise. It can be easily shown that under such circumstances the optimal inference about the causes in a cue combination experiment combines the measurements linearly. Although this is only a special case, this makes computations much easier. While Gaussian noise may not be the appropriate model for many real world processes, many laboratory experiments in which the researcher can control the perturbations of the stimuli use this paradigm for simplicity. The optimal linear cue weights by which the individual estimates have to be weighted to obtain the best combined estimate can be shown to be proportional to the inverse variances of their respective cues:

$$w_a = \frac{\frac{1}{\sigma_a^2}}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_v^2}} \qquad w_v = \frac{\frac{1}{\sigma_v^2}}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_v^2}} \tag{3}$$

Intuitively, this is a satisfying solution in that it means that cues that have a higher associated uncertainty are relied upon less, while cues that have a lower uncertainty influence the final result more. Importantly, the variance of the combined estimate is smaller than the individual variances and can be shown to be:

$$\sigma_{av}^2 = \frac{\sigma_a^2 \sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{4}$$

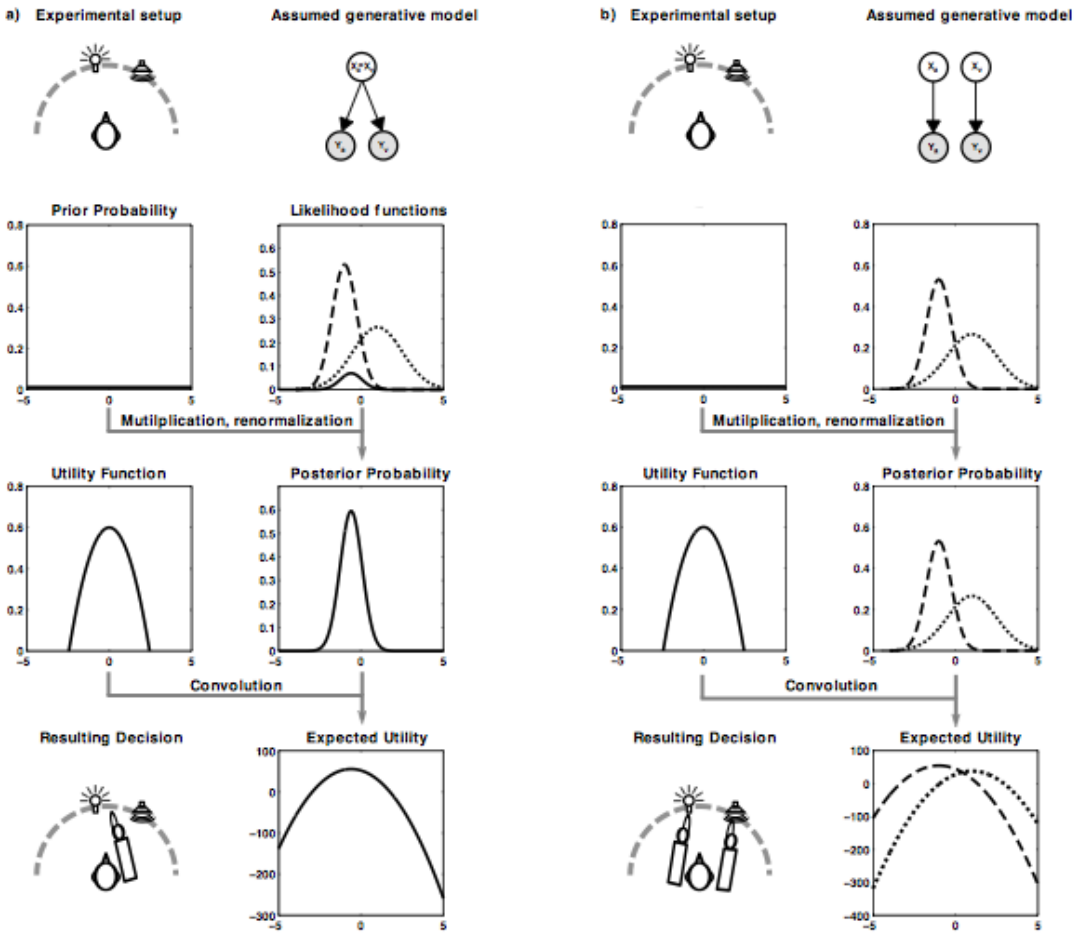

Figure 2: Graphical depiction of the steps needed to reach a decision according to two Bayesian models in an audio-visual localization task. a) Decision reached by a model subject when sensing a light flash and a brief tone when using a model assuming that both stimuli come from a common source. b) Decision reached by a model subject when sensing the same light flash and a brief tone when using a model assuming that the two stimuli come from two distinct sources.

Non-linear cue combination is in principle nothing special, as the Bayesian formulation in equations (1) and (2) make it clear, that the full likelihood functions have to be considered in general to compute the posterior at each value of X. Nevertheless, because of the computational simplicity, the vast majority of cue combination experiments have used stimuli such that linear cue combination can be

used. Nonlinear cue combination can arise in different ways. The simplest case is that one of the component likelihood functions is non-Gaussian. It may also be that the individual likelihood functions are Gaussian, but that the task at hand does not require estimating an additional scene parameter, which nevertheless influences the measurements. This leads to an estimate that discounts the uninteresting variable and the associated marginalization procedure, which collapses the probability in the uninteresting variable onto the distributions of the cues, leads to the computation of an integral for the combined likelihood function that usually results in a highly non-Gaussian shape (e.g. Saunders, Knill 2001). A similar situation arises, when the likelihood function is modeled as a mixture distribution in order to accommodate different scene configurations (Knill 2003).

The prior distribution P(X) has a particularly interesting role in many modeling situations. The physical world is characterized by a large number of structural regularities, which manifest themselves in certain states of scene parameters having much higher probability than other states. A good example for the impact of the prior distribution comes from a study by Weiss, Simoncelli, and Adelson (2002) in which human motion perception was considered. It is well known from the area of computer vision that the direction and the speed of motion is ambiguous when inferred from local image motion, that is, a single measurement of a single motion cue gives rise to an infinite number of possible true motions in the scene. This relationship is expressed in a constraint equation, which linearly relates the velocity in horizontal direction with the velocity in vertical direction given a local velocity measurement. The likelihood function expresses this linear relationship and additionally reflects the fact, that the uncertainty about the image motion increases when the contrast in the image display is reduced. Importantly, Weiss, Simoncelli, and Adelson (2002) noticed, that a wide variety of perceptual phenomena in motion perception and visual illusions with moving stimuli could be explained by assuming that humans use a prior distribution over motion velocities that makes zero velocity and low velocities more likely. There is considerable literature on characterizing prior distributions both by finding those distributions that best explain psychophysical performance (e.g. Weiss et al. 2002, Stocker & Simoncelli 2006) as well as directly measuring distributions of physical states of the world (e.g. Geisler et al. 2001). These studies have shown that the prior distribution is often non-Gaussian in real world settings and that picking a particular prior distribution is the key in explaining observed behavior.

After the above computations have been executed, the ideal observer has obtained a posterior probability distribution over the unknown scene parameters that are of interest, as the position x of the audio-visual source in the localization example.
Given that perceptually we experience a single state of the world instead of a posterior distribution over possible scene parameters, the posterior distribution has to be utilized to come to a single estimate corresponding to the perceived scene. Similarly, when an action is required such as reaching for a point in the scene, a single direction of motion of the hand has to be obtained from the entire probability distribution. The system therefore has to collapse the posterior probability distribution over all possible states of the relevant scene parameter to a single value, which can be done in different ways, depending on the used criterion. The literature on optimal decision making and utility functions is vast not only in the areas of cognitive science and statistics but especially in economics (Neumann & Morgenstern, 1944) and can be tracked back to Bernoulli (1738). Bayesian decision theory deals with how to come to such a single decision given a distribution over X and a loss function $L(\hat{x})$, which measures the penalty of deciding that the value of X is $\hat{x}$ when the true value is

X. Here, the relevant distribution is the posterior distribution over the scene parameters. The expected loss, which quantifies how much we expect to lose with a decision given the uncertainty in the variable X, can be expressed as follows in the case of the audio-visual localization task:

$$L(\hat{X}|Y_a, Y_v) = \int L(\hat{X}, X) P(X|Y_a, Y_v) dX \tag{5}$$

By calculating this point-by-point product of the loss function with the posterior and then integrating over all possible positions x, one obtains the expected loss as a function of $\hat{X}$. The optimal decision is then to choose that value $\hat{X}$ that minimizes the expected loss. While a wide variety of cost functions are possible, only a few are commonly used in practice, such as a cost function that assigns a one to the decision corresponding to the true value and a zero otherwise, or linear costs, or quadratic costs. In the special case of a Gaussian distribution, these cost functions give the same result, corresponding to choosing the maximum a posteriori estimate (MAP), which is the value of highest posterior probability. This sequence of computations is depicted in the figure 2a) for the audio-visual localization task.

Bayesian ideal observers have been designed for a large variety of tasks involving many different environmental properties and cues and they can accommodate the computational principles mentioned in the above section such as cue combination (figure 1c), cue integration (figure 1c), explaining away (figure 1d, e). In principle, there is no limit to the number of different variables that make up the generative model, although exact computations may have to be replaced by approximations due to complexity. As an example, consider the case in which the observer needs to infer the identity of an object but obtains image measurements that are dependent on the position of the illumination source, the orientation of the object towards the observer, the specularity of the surface of the object, and many additional variables describing the scene. In such a case, the inferential process needs to compute the posterior probability of the potential target objects by discounting the non relevant variables such as illumination direction in a probabilistically optimal way. Such marginalization calculations can be very involved, but it has been shown, that there are situations in which human subjects take such uncertainties into account (Kersten, 1999). For further examples see the review by Kersten and Yuille (2003) that presents several other influence diagrams representing different dependencies between observed and non-observed variables that may be of interest or need to be discounted for the inference task at hand.

Bayesian models have the advantage compared to other methods that they can accommodate computations with uncertainties, because these are explicitly represented as probability distributions. Indeed, incorporating sequential decisions and explicit costs leads to a representation of all these tasks as Partially Observable Markov Decision Processes (POMDPs), which are very versatile and general, but have the disadvantage of being computationally intractable for most realistic problems. An important further advantage of Bayesian models is that they allow computing quantitatively how well different models are applicable to data, at least in principle. The idea is to assign a posterior probability to different models given the observed data (see e.g. MacKay, 2003; Bishop, 2006). In general, if arbitrary models are considered, this calculation has to take into account each model's complexity, which may result in intricate computations. But recent work has started to apply Bayesian methods that explain the selection of the appropriate model (Koerding et al., 2007; Sato et al., 2007) under different stimulus conditions, by computing how likely the observed data is under different models. Models for the task of model

selection in multimodal object perception have been proposed by Sato et al. (2007) and Beierholm et al. (2008) for audio-visual cue integration tasks.

The popularity of the Bayesian approach is supported by the fact that it allows formulating a quantitative measure of optimality against which human performance can be compared. Ideal observer models that match behavioral performance of humans and primates have been found for a large number of such tasks. But, importantly, humans are never as efficient as a normative optimal inference machine i.e. they are not only limited by the uncertainties of the physical stimuli, but may also be limited by uncertainties in the coding and representation process, have limited memory, and the computations themselves may have an associated cost. This means, that generally it is not the case that human performance is optimal as that of an ideal observer, but it simply means that it is tested whether the brain executes computations by taking uncertainties into account. Nevertheless, it is worth stressing the importance of this basic fact and its historic relevance, as previous research e.g. by Kahneman and Tversky (2000) showed, that in many cognitive tasks, humans do not correctly take the respective uncertainties into account. This means that subjects in cognitive decision tasks fail to minimize expected loss but in mathematically equivalent movement tasks often choose nearly optimal (Trommershäuser et al., 2008).

Unfortunately, the normative formulation also has a number of limitations. First, while it is important to formulate a model of optimal behavior, such models are always bound by their respective assumptions. As an example, most cue integration studies assume that the underlying probability distributions are implicitly known by the subjects, but it is not clear, how these distributions are learned. Accordingly, many of such studies do not report and are not interested in the actual learning process during the cue integration experiments. An interesting aspect of this problem is, that the often referred to concept of internal noise can actually veil a number of possible sources of uncertainty. As one example, if a human subject used a likelihood function that is not the correct one describing the generative process of the data, performance on an inference task may be suboptimal and the experimenter might interpret this as evidence for internal noise.

An optimality formulation may seem to be straightforward to come about for a range of experiments, in which the experimenter can fully control impoverished and constrained stimuli. But in Bayesian terms it is often possible to model the respective data in several ways. Consider a linear cue integration paradigm. The assumption in such experiments is, that the system knows the respective uncertainties in the underlying Gaussian distributions. But presumably, in experiments where these uncertainties have to be learned, these uncertainties need to be estimated over trials. In a Bayesian model, it would be necessary to model the task with explicitly representing the uncertainties over the parameters, i.e. one could model each cue in terms of a Gaussian distribution with unknown mean and unknown variance and introduce a distribution over the two unknown parameters, such as an Normal-scaled inverse gamma distribution. But many alternatives exist, and it is currently not clear, how to decide which model best describes human and animal performance.

Furthermore, it may be straightforward to write down a generative model for a specific task, but the inversion of the model, which is required for the inference process, may be a rather daunting task that requires extensive approximations. For example, in model selection it is often difficult to calculate the posterior distribution of the data given different models and usually approximation techniques have to be

used. While it has been suggested that the brain could carry out computations such as Markov chain Monte Carlo (MCMC) (Hinton, Sejnowski, 1986; Hoyer, Hyvärinen, 2003) and recent work has used such algorithms to model subjects' behavior in cognitive tasks (Sanborn et al., 2006), this remains an open question that will require further research.

3.4.2 Intermediate models of multimodal object perception

There is considerable literature that has been interested in modeling multisensory object perception at an intermediate level of description. This discussion has been partly motivated by attempting to map algorithms from the domain of machine perception and especially computer vision to human visual perception. It has also been motivated by the hypothesis that individual areas along the cortical hierarchy compute separate perceptual estimates. A central topic in this literature has been that of computational modules. If one assumes that distinct cortical regions are responsible for specific computations involved in the extraction of distinct features, it is natural to ask how the individual results of computations can be combined to obtain a globally optimal estimate of a scene parameter. While such models have also employed Bayesian techniques, the emphasis here is more on how an optimal estimate of a scene parameter may be obtained if specific separate computations by individual modules have to be combined to a globally optimal estimate.

Yuille and Bülthoff (1996) consider the case where the posterior distribution of a scene variable given two cues is proportional to the product of the two posterior distributions given the individual cues, i.e. $P(X \mid Y_1,Y_2) = c\ P(X \mid Y_1)\ P(X \mid Y_2)$. In terms of the ideal observer analysis, note that in cue-combination paradigms with independent Gaussian noise an estimate for the unknown scene property can be computed by obtaining the mean and variance of the posterior for each individual cue separately. These separate estimates can be weighted linearly to obtain the mean and variance of the maximum a posteriori combined estimate. The question therefore arises, whether these types of calculations are carried out separately first before combination or whether these computations are always carried out jointly.

Consider an example by Ernst and Bülthoff (2004) in which a human subject estimates the location of its hand knocking on a surface. There are visual, auditory, and proprioceptive cues to this location, but the signals are represented in different coordinate frames and units. The auditory and visual signals have to be combined with the postural signals of the head into a coordinate frame that allows combination with the proprioceptive signal from the knocking hand. Again, the question arises whether the auditory and visual signals are first separately transformed into different coordinate systems followed by a combination with the proprioceptive signal. The alternative is that these computations are done by a single calculation in which all separate signals are directly combined to an optimal estimate. In this context, weak fusion refers to independent processing of individual cues followed by a linear combination of these estimates, while strong fusion refers to computations in which the assumed cue processing modules interact nonlinearly. Extensions to these types of models have been proposed in the literature (Yuille, Bülthoff, 1996; Landy et al., 1995).

While it is important to ask how the brain actually carries out such computations, it should be noted that the ideal-observer framework can accommodate these views in a straightforward way by using the Bayesian network formalism. Considering the above example, if the sensory noise in the two cues is conditionally independent, the

ideal observer model expresses this as a factorized distribution, i.e. $P(Y_1|X)P(Y_2|X)$. If instead the noise is not independent, the full joint probability distribution $P(Y_1, Y_2|X)$ needs to be specified. Thus, weak and strong fusion can be matched to factorizations underlying the generative model of the data. Similarly, nonlinearities can be introduced by marginalization computations when discounting scene variables that are not of interest in the task but determine the values of the observed cues. Finally, internal computations such as sensorimotor transformations can be incorporated in Bayesian ideal observers by choosing appropriate probability distributions in the corresponding likelihoods of cues or explicitly as an additional intermediate random variable.

3.4.3 Neural models implementing multimodal perception

There is a line of research that has pursued the modeling of multimodal perception at the level of individual neurons and networks of neurons. This research has often approached this goal by looking at response properties of neurons in specific brain regions and trying to obtain similar behavior starting from neuronal dynamics and learning rules. Work in this category has e.g. considered the multisensory responses of single neurons in the superior colliculus, which only fire when signals from different modalities are close in time and space (Meredith, Stein, 1986; Wallace, Wilkinson, Stein, 1996). Some neurons also show an enhanced response to a multisensory stimulation that is larger than the linear sum of the responses to the individual signals, often called Multisensory Enhancement (MSE) (e.g. Alvarado et al. 2007). Models by Anastasio and Patton (Anastasio, Patton, 2003; Patton, Anastasio, 2003) try to explain the development of MSE through a two-stage unsupervised learning mechanism, which includes cortical modulation of the sensory inputs. Rowland, Standord and Stein (2007) developed a model of the processes in a single multisensory neuron that also shows MSE. It is motivated by the characteristics of NMDA and AMPAR receptors in neurons as well as inspired by the fact that superior colliculus neurons obtain their inputs from ‘higher’ cortical areas as well as from sensory areas. These two models used artificial inputs and looked at a limited set of processes, but they nevertheless show that some computations for a possible implementation of Bayesian principles in the brain could be accomplished with basic neuronal mechanisms (Anastasio et al., 2000).

The success of Bayesian models in accounting for a large number of experimental data has lead to the more general idea that neural computations must be inherently probabilistic and maybe even explicitly represent probability distributions. Accordingly, a different line of theoretical research has started investigating models in which neuronal activity represents and computes with probability distributions. In principle, the reviewed studies have demonstrated that human and primate behavior is taking the relative uncertainties of the cues into account. This does not imply that the computations carried out in the brain have to be Bayes optimal, but at least this implies, that the relative uncertainties have to be represented somehow. Nevertheless, major efforts are currently directed at proposing neural representations that explicitly encode uncertainties in sensory representations together with suggestions how computations such as cue integration, maximum-likelihood estimation, and Bayes optimal model selection could be carried out.

Coding schemes have been proposed in which the activity of populations of neurons directly represents probability distributions. We briefly review the key ideas of such coding models and refer the interested reader to the review by Knill and Pouget (2004) and the book by Doya, Ishii, Pouget, Rao (2007) for more detailed

descriptions. The starting point of the proposed coding schemes is a new interpretation of the probabilistic nature of neuronal responses. Responses of neurons to the repeated presentations of a stimulus to which they are tuned show considerable variability from trial to trial, i.e. their activity has been characterized as being noisy. This response variability has now been thought instead to be inherently related to the probabilistic nature of the stimuli itself. Several models have proposed different ways of how to map this idea to different random variables, which can be continuous or discrete.

A straightforward model is to have a neuron's firing rate be some function of the probability of an event. Gold and Shadlen (2001) have proposed such a model based on neurophysiological data obtained from LIP neurons during a motion discrimination tasks, in which monkeys had to decide which of two possible motion directions a random dot kinematogram had. Decisions taken by the monkey could be predicted by a decision variable, which was calculated as the logarithm of the difference between firing rates of neurons. By relating this decision variable to the probability of motion direction, this result was taken as evidence for a direct encoding of a function of the posterior probability of motion direction given sensory stimuli. Deneve (2005) extended this model to spiking neurons, in which each neuron represents the ratio of the logarithms of the probability of two preferred states and the neuron only fires a spike, if the difference between a prediction and a current observation is exceeded.

A different direction in modeling encodings of probability distributions is motivated by the fact that large populations of neurons jointly represent sensory variables. Individual neurons in such populations are distinct from each other by their respective tuning curve, i.e. by the mean firing rate as a function of the considered stimulus value. As an example, V1 simple cells' activities vary with the orientation of a stimulus and individual cells have different preferred orientations. There exist two closely related representations of probability distributions over scene parameters referred to as 'convolution encoding' and 'convolution decoding'. Both proposals are based on the mathematical idea of a change of bases. In essence, one can approximately synthesize a function or probability distribution by a linear sum of individual prototype-functions, so called basis functions. In order to find the linear coefficients for the sum of basis functions to return the original function, this has to be projected down onto the new basis set.

In both convolution encoding (Zemel at al., 1998) and convolution decoding (Anderson et al., 1994) a neuron's activity is related to an approximation of the true posterior P(X|Y) using basis functions, which are often chosen to be Gaussians. While in convolution decoding neuronal firing rates represent the coefficients that are multiplied with their associated basis functions to give their respective contributions to the approximate posterior, in convolution encoding the posterior is projected down onto individual neuron's tuning curves. Because neuronal activity is modeled as being stochastic, the additional Poisson noise ends up having different influences on the resulting probabilistic encoding and decoding schemes. One of the well known difficulties with these coding schemes is that highly peaked distributions corresponding to small amounts of uncertainty get much broader. Nevertheless, considerable work has investigated, how computations necessary for multimodal inference could be carried out on the basis of these coding schemes.

A recent approach to the representation of probability distributions and importantly also to inference computations with these representations in neurons comes from the

work of Ma, Beck, Latham, and Pouget (2006). The fundamental idea of this probabilistic population code is that the activity of a population of neurons will at some point be decoded to obtain a posterior distribution over the stimulus dimension X given this neuronal population activity r. Given that the neuronal response is r, this can be formulated as inference of the stimulus value X based on Bayes theorem:

$$P(X|r) = \frac{P(r|X)P(X)}{P(r)} \tag{6}$$

The common way of quantifying neuronal population activity is by tuning curves describing the activity of individual neurons to the stimulus value. Given the near Poisson variability across repeated trials, this mean response value given by the tuning curves corresponds to the variance of the activity, given that mean and variance have the same value for the Poisson distribution. Under certain conditions, such as independent Poisson variability across neurons, the posterior P(Xlr) converges to a Gaussian distribution, whose mean is close to the stimulus at which the population activity has its peak, while the variance in the posterior is inversely proportional to the amplitude of the activity hill. So, a stimulus with high uncertainty corresponds to a small height of the hill of activity, whereas low uncertainty corresponds to large amplitude.

In a situation requiring cue integration, the brain needs to compute an estimate from two such hills of activity and reach a result that corresponds to equation (2). Ma et al. (2006) showed that this point by point multiplication of probability distributions could be carried out simply by adding the two neuronal populations in the probabilistic population code framework. Furthermore they showed that only Poisson-like neuronal variability allows for such a computationally straightforward scheme, which can easily be implemented at the neuronal level. Thus, the interesting result from this model is that it provides a new interpretation of the near-Poisson variability of neuronal activities as a basis for computing with uncertainties. It is important to note however, that these proposals encode those types of uncertainties that are readily given in the stimulus such as the uncertainty of a visual stimulus by virtue of its contrast. This is different from situations in which cues have an associated uncertainty that is in turn inferred from the reliability of a cue's statistic on a very fast timescale (Triesch, Ballard, Jacobs, 2002) or when the reliabilities are learned over longer timescales (Fine, Jacobs, 1999).

3.5 Open questions

3.5.1 How are model parameters learned?

If the brain is capable of computing with uncertainties in multisensory signals according to Bayesian inference, then it is of interest to understand where the parameters of the probability distributions in the generative model describing the relationship between the hidden causes and the observations come from. A variety of techniques in machine learning allow the computation of the parameters in a given generative model, if the model is somehow specified a priori (e.g. Bishop, 2006; MacKay, 2003). Maximum likelihood techniques allow setting the parameters such that the probability of observed data is maximized under the given model. This can also be achieved, if some variables in the fully specified model are unobserved or latent by using the Expectation-maximization algorithm (EM) or related techniques.

The EM algorithm iteratively alternates between an expectation (E) step, which computes an expectation of the likelihood given the current guesses for the latent

variables and a maximization (M) step, which maximizes the expected likelihood found in the E step by adjusting the parameters. Repeating this procedure can be shown to increase the likelihood of the data given the parameters, but may only find a local minimum in the likelihood. But how the brain can accomplish such computations is still unknown.

In case of the audio-visual localization task, the brain must somehow find the parameters of the Gaussian distributions that describe the uncertainty in the auditory and the visual location measurements. In a laboratory version of this task the uncertainties can be varied e.g. by changing the contrast of the visual stimulus or the loudness of the sound. This means, that the system needs to compute with the uncertainties associated with the respective contrast and loudness. Similarly, given that acuity is dependent on the eccentricity, these uncertainties are also dependent on the relative position of the stimuli with respect to the auditory and visual foveae. Furthermore, for ecologically valid contexts, these distributions are often dependent on many more parameters that the system needs to take into account.

A further complication under natural conditions is that the relative reliabilities of cues may change over time and could also be context dependent. Similarly, individual cues could acquire biases over time. Studies by Jacobs and Fine (1999) and Ernst et al. (2000) have shown that humans are able to adjust the relative cue reliabilities and biases when multiple cues are available and one of the individual cues is perturbed artificially. While these studies required subjects to learn to adjust the relative cue weights over timescales of hours and days, a study by Triesch et al. (2002) showed that such adaptation processes could also be observed on the timescale of seconds. In their experiment subjects had to track individual objects in a display containing several distractors and detect the target object after an occlusion. Objects were distinguishable by the three cues of color, shape, and size. While being occluded, some of the features defining the object identity were changed. By adapting to the relative reliabilities of individual features, subjects reweighed the cues depending on their respective reliabilities on the time-scale of one second.

3.5.2 How are the likelihood functions learned?

Similarly, but even more difficult to answer with current knowledge is the question of how the brain is able to learn the likelihood functions describing the relationship between individual scene parameters. One of the main tasks of the modeler is often to come up with the correct likelihood function but it is not clear, how the primate brain accomplishes this feat. As an example, how does the brain know that texture information about the slant of a planar surface is less reliable at low slants than at high slants (Knill, Saunders, 2003)? And how does the brain know the uncertainty in orientation of a visual stimulus as a function of its contrast?

3.5.3 Where does the prior come from?

If the system is able to use prior distributions over scene parameters it must have stored such knowledge. Although it is conceivable that some such prior distributions are encoded genetically, there is also evidence that they can change due to learning. Adams et al. (2004) used the well known human prior in the judgment of shape from shading which prefers interpretations of ambiguous two dimensional shaded image regions as being lit from above. Subjects' shape judgments changed over time with visuo-haptic training, in which the haptic stimuli were consistent with a light source shifted by 30 degrees. Moreover, a different subsequent lightness judgment task revealed, that subjects transferred the shifted prior on illumination direction to new

tasks. This suggests that subjects can learn to adjust their prior on illumination direction from visual haptic stimuli. Similarly, Knill (2007) showed that humans learn to adapt their prior, which favors the interpretation of ellipses as slanted circles. The data was modeled qualitatively with a two-step Bayesian model in which a first process uses the current prior over elliptical shapes to estimate the aspect ratio of the currently presented ellipse. A second process then updates the current prior using the estimated shape of the viewed ellipse.

Additional complexities in learning priors can be found in considering the case of motion perception from local estimates of image motion. While the work by Weiss, Simoncelli, Adelson, (1999) shows that human experimental data can be well explained by assuming a prior distribution that favors velocities near zero, it is not clear how such a prior is learned, if at all. One could hypothesize that such a prior is learned only on the basis of the statistics of image motion but it is also conceivable that it is learned after image segmentation and by compensating for the image motion induced by selfmotion of the observer, i.e. that the prior is learned on the basis of the estimated motion distributions of objects in the world.

While the vast majority of work on the origin of prior distributions has focused on characterizing the statistics of properties of the natural environment such as luminance, contrast, disparity, sound wave amplitude, these investigations have almost exclusively characterized these distributions without taking into account the active exploration of the environment. But the actual input to the sensory system is without doubt dependent on the ongoing behavior, which often samples the environment in structured ways. Rothkopf and Ballard (2009) showed that the usage of the sensory system crucially influences the statistics of the sensory input to the visual system in a task dependent manner and that therefore the input to the visual system significantly depends on behavior. Thus, when characterizing the prior distributions of scene variables, which are considered to be the input to the sensory system, it is important to take the organism's active behavior into account.

Indeed, a further study by Rothkopf, Weisswange, and Triesch (2009) explored how the sensory apparatus itself and its active use during behavior determine the statistics of the input to the visual system. A virtual human agent was simulated navigating through a wooded environment under full control of its gaze allocation during walking. Independent causes for the images obtained during navigation were learned across the visual field with algorithms that have been shown to extract computationally useful representations similar to those encountered in the primary visual cortex of the mammalian brain. The distributions of properties of the learned simple cell like units were in good agreement with a wealth of data on the visual system including the oblique effect, the meridional effect, properties of neurons in the macaque visual cortex, and functional Magnetic Resonance Imaging (fMRI) data on orientation selectivity in humans and monkeys. Crucially, this was only the case if gaze was allocated overwhelmingly in the direction of locomotion, as is the case in natural human walking. But when gaze was directed mostly to a point on the ground plane, the distributions of properties of the learned simple cells differed significantly from those consistent with the empirical findings in humans and primates.

3.5.4 How does the brain learn the appropriate generative model, if at all?
This question extends the above points in that the full generative model that is required for inferring unobserved scene parameters obviously includes the structure of the variables describing the scene and all available cues. For example, one needs

to know that shading is dependent on the illumination direction but independent from self-motion. With the very large number of multisensory variables that primates are capable of extracting from their environment this is a daunting task. How such independencies are acquired, is still unknown. Again, the field of machine learning offers algorithms for learning the structure of a generative model given observed data. In its most general form this problem is NP hard and no analytical solution is available. Under certain restrictions on the types of possible models, i.e. on the way individual variables are independent of others, there are algorithms that allow for a more efficient calculation of the probability of observing the given data under a particular model, thus allowing selecting the best fitting model. Approximate structure learning in Bayesian Network models is an area of active research.

It is also important to note that some experimental data does not conclusively answer the question of whether subjects use the correct likelihood functions in the first place. As an example, a study by Oruc, Maloney, Landy (2003) showed that there are significant differences between individual subjects that can be explained by the assumption that they used different cue integration strategies. The study looked at human estimation of slant from linear perspective and texture gradient cues. The idea was that the noise in the two cues may be correlated, as the same visual neurons were involved in representing both cues. The behavioral data was best explained by the assumption that some subjects used a cue combination rule consistent with independent noise in the cues while others assumed correlated noise. This study also points towards the problem, that it is difficult to assess, whether a subject uses exactly a specific generative model in an inference task, as errors due to an approximate or simply wrong likelihood model could look like noise.

A further question that arises from these studies is whether human subjects can in principle learn arbitrary models. Two studies by Ernst (2007) and by Michel and Jacobs (2007) investigated this question. The idea was to involve human participants in a learning study in which they were exposed to artificial cue congruencies that are not encountered in natural environments and to test whether these cues were combined in a subsequent cue combination task. The former study used visual and haptic stimuli while the latter used several visual and auditory cues. Interestingly, while the former study found significant effects of learning and concluded that humans can learn arbitrary signals across modalities, the second study came to the conclusion, that humans can learn the parameters of the implicit generative models, but not novel structures such as causal dependencies between usually independent variables.

A different approach to this problem was recently proposed by Weisswange, Rothkopf, Rodemann, Triesch (2009). The task in this case is again an audio visual orienting task, in which the auditory and visual sources may coincide or may be originating from different sources. A learner is rewarded for orienting towards the true location of the object on repeated trials, in which the two sensory stimuli can appear at different positions. This study used a basic reinforcement learner, i.e. a system that learns to do optimal control on the basis of exploring the environmental dynamics and obtaining feedback from the environment about the success of its actions. The optimal control in this case is to orient toward the true position of the light source. Interestingly, after learning, the agent was able to orient towards the most likely position of the target by taking the relative reliabilities of the auditory and visual cues into account for the integration. Furthermore, when the two sources are further apart, the agent learned to not integrate the two cues, but to rely only on the

less uncertain visual measurement. Thus, although this study does not conclusively show that reinforcement learning alone is the basis for learning cue integration and causal inference, it provides a mechanism by which such abilities could be learned from interacting with the environment.

3.5.5 How are different models combined?

Recent work on causal inference in multisensory perception (Koerding et al., 2007) has suggested that humans have multiple generative models that describe data for different scene configurations. In their experiments, subjects are modeled as computing the likelihood of an observed flash and a tone under the assumption that these came from the same source (figure 2a) and under the assumption that they came from two different sources (figure 2b). While under such laboratory situations it is possible to construct very controlled stimuli that are compatible only with few potential scene configurations, in ecologically valid situations the number of potential models that have to be compared grows quickly. How the brain may choose the appropriate models for comparison or how the brain may learn the appropriate models is an open question.

Furthermore, model selection may not be the optimal way of combining different models. Indeed, the Bayesian method of combining the estimates of different models is again to weigh the different estimates according to the uncertainty associated with the data given each model. To make things even more complicated, it is in principle also possible to assume that the parameters of individual models change over time. This requires the system to obtain additional data over trials in order to estimate the changes in the parameters. Under such circumstances it is possible to develop models that use a strategy of probability matching. Currently, there is conflicting evidence as to what primates do under such circumstances.

3.5.6 Does the brain represent full probability distributions or implicit measures of the uncertainties?

The jury on how the brain computes with uncertainties is still out. Some empirical studies have tried shedding more light on this issue. Work by Körding and Wolpert (2004) on sensorimotor learning used a task in which human subjects had to reach to a visual target with their index finger in a virtual reality setup. Visual feedback about the finger's position was given only through a set of light dots. This allowed introducing a conflict that displaced the center of the light cloud relative to the true position of the index finger. The lateral shifts on single trials were obtained as samples from different complex probability distributions in different learning experiments. Some subjects were exposed to shifts drawn from a Gaussian distribution, while others were exposed to shifts coming from a mixture of two or three Gaussian distributions. The overall result of the experiments was that subjects took the uncertainties in the visual feedback and of the distribution of applied shifts into account in a way that was compatible with a Bayesian strategy. But subjects did not behave optimally when a complex distribution such as the mixture of three Gaussians was used, at least within the allotted number of approximately 2000 trials given by the experimenters. This suggested to the authors, that there are limits to the representation or computation of arbitrary probability distributions describing the uncertainties in a task.

At the neuronal level, the models of probabilistic population codes described above propose different ways how populations of neurons could encode full probability distributions or functions of them. Further research will need to develop experiments

that can disambiguate different proposals. It should also be noted, that there is evidence that the brain may have developed a number of different codes depending on the required task or stimulus dimension. Work by Ahissar and colleagues (see Knutsen & Ahissar, 2008 for a review) has demonstrated that tactile encoding of object location in rodents uses timing and intensity of firing separately for horizontal and radial coordinates. Accordingly, it may very well be that the brain uses different solutions to the problem of representing uncertainties.

3.5.7 How ideal is human learning?

The vast majority of cue combination experiments report averaged behavioral data and experiments that include a learning component rarely report individual learning curves. Indeed, the performance of humans is far from the Bayesian ideal observer in terms of utilizing the available information acquired during learning experiments, if one applies the true generative model from which the stimuli were generated. The problem here is, that it is not exactly known, what model is actually used by the brain, or whether the brain maintains several models in parallel. These and similar observations from the animal learning literature have led to Bayesian learning models in the areas of cognitive tasks (Sanborn, Griffiths, Navarro, 2006) and animal conditioning (Daw, Courville, 2008) that explicitly model how individual decisions on a trial may be modeled by assuming that the brain does inference using sequential Monte Carlo sampling. The idea is that, instead of maintaining a full belief distribution, at each decision one sample from one of the current hypotheses guides the decision. Further research will be needed to clarify what methods the primate brain actually implements.

3.5.8 How do laboratory experiments transfer to natural tasks?

In ecologically valid situations there is an abundance of ‘cues’, e.g. there are at least a dozen depth cues in natural, rich scenes. While cue integration experiments have demonstrated that two or three such depth cues are integrated into a percept that takes the inverse variances into account, it is not known whether this strategy also transfers to the cases of rich scenes, in which a few cues are much more reliable than others. Are really all cues always integrated by default? If computations do not have an intrinsic associated cost or, equivalently, there are no restrictions on the durations for the computations to be terminated, it is of course of considerable advantage to have a multitude of different cues. Under such conditions, the respective biases and uncertainties can be corrected for by recalibrating those cues that disagree most with results that have been successful in the recent past. This is exactly the approach of Triesch and v.d. Malsburg (2001). In their democratic cue integration model, a multitude of individual sensory cues are brought into agreement and each cue adapts towards the agreed upon result.

3.5.9 The role of approximations

While ideal observer models often use techniques for approximate Bayesian inference, because there are no analytic solutions available, the role of approximations with regard to the brain has been underestimated. It is currently unknown, how the brain accomplishes computations that are in accordance with Bayesian inference. But it may very well be, that the brain is confronted with similar difficulties as researchers in machine learning in carrying out such computations, and that the approximations that are used to solve the inference problems are of central importance. Similarly, the involved computations themselves may have different effects on the associated uncertainties. Thus, what are the approximations that the brain uses? Can these approximations be expressed in terms of cost functions, i.e.

are there intrinsic behavioral costs that influence what type of approximations are adequate?

One particular uncertainty introduced by computations carried out by the brain itself stems from sensorimotor transformations. Consider again the earlier example of a task requiring the system to integrate two position estimates that are represented in different coordinate frames, such as an eye-centered and a head-centered reference frame. In order to plan and execute a reach, such information has to be remapped into a common coordinate frame, which has to take into account the noisy proprioceptive estimates of the involved joint angles. These computations introduce additional coordinate transformation uncertainty. A study by Schlicht and Schrater (2007) investigated whether human reaching movements take these uncertainties into account and found evidence that subjects adjusted their grip aperture reflecting the uncertainties in the visual stimuli and the uncertainties due to coordinate transformations. The behavioral human data was well accounted for by a Bayesian model built on the assumption that internal spatial representations used an eye centered reference frame for this task. From a theoretical point of view, there are many more instances of uncertainties that are the result of computations and further research may reveal additional insight into whether the nervous system takes these into account. Considerable further research will have to determine the experimental and analytical tools required to address these questions.

3.6 Conclusion

Models of multisensory object perception have been developed over the last decades that have had great success at explaining a multitude of behavioral data. The main reason for this success has been to recur to models that not only explicitly represent a sensory estimate, but also the uncertainties associated with them. Bayesian techniques have been especially successful, as they allow for the explicit representation of uncertainties of model parameters. Nevertheless, a wealth of open questions remains to be answered, ranging from how humans and primates learn parameters and models connecting sensory stimuli and world properties, to how the neuronal substrate may compute with uncertainties.

References

Adams, W.J., Graf, E.W., Ernst, M.O. (2004) Experience can change the 'light-from-above' prior. Nat Neurosci 7(10):1057-8

Alais, D. and Burr, D. (2004): The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14(3):257–62

Alvarado, J.C., Vaughan, J.W., Stanford, T.R., Stein, B.E. (2007) Multisensory versus unisensory integration: contrasting modes in the superior colliculus. J Neurophysiol 97(5): 3193-205

Anastasio, T.J., Patton, P.E. (2003) A two-stage unsupervised learning algorithm reproduces multisensory enhancement in a neural network model of the corticotectal system. J Neurosci 23(17): 6713-27

Anastasio, T.J., Patton, P.E., Belkacem-Boussaid, K. (2000) Using Bayes' rule to model multisensory enhancement in the superior colliculus. Neural Comput 12(5): 1165-87

Anderson, C. H., Van Essen, D. C. (1994). Neurobiological computational systems. In J. M. Zureda, R. J. Marks, & C. J. Robinson (Eds.), Computational Intelligence Imitating Life (pp. 213–222). New York: IEEE Press

Atkins, J.E., Fiser, J., Jacobs, R.A. (2001) Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. Vision Res 41(4):449-61

Battaglia, P.W., Jacobs, R.A., Aslin, R.N. (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20(7):1391-7

Battaglia, P. W., Schrater, P., & Kersten, D. (2005). Auxiliary object knowledge influences visually-guided interception behavior. Proceedings of the 2nd symposium on applied perception in graphics and visualization, ACM International Conference Proceeding Series, pp. 145 - 152.

Bernoulli, Daniel; Originally published in 1738; translated by Dr. Lousie Sommer. (January 1954). "Exposition of a New Theory on the Measurement of Risk". Econometrica 22 (1): 22–36

Beierholm, U., Kording, K., Shams, L., Ma, W. J. (2008) Comparing Bayesian models for multisensory cue combination without mandatory integration. Advances in Neural Information Processing Systems 20, 81-88. MIT Press, Cambridge, MA.

Bishop, C.M. (2006) Pattern recognition and machine learning. Springer-Verlag, New York

Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., King, A.J. (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17(9):2172-89

Bruce, C., Desimone, R., Gross, C.G. (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46(2):369-84

Bülthoff, H.H. and Mallot, H.A. (1988) Integration of depth modules: stereo and shading. J. Opt. Soc. Am. A 5, 1749 – 1758

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S. (1997) Activation of auditory cortex during silent lipreading. Science 276(5312):593-6

Clark, J.J. and Yuille, A.L. (1990) Data Fusion for Sensory Information Processing Systems. Kluwer

Daw, N.D., Courville, A.C. (2008) The pigeon as particle filter. in Proceedings of NIPS 2008

Deneve, S. (2005) Bayesian inferences in spiking neurons. in Proceedings of NIPS 2005: 353-360

Doya, K., Ishii, S., Pouget, A., Rao, R.P.N. (2007) The Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press, Cambridge

Ernst, M.O. (2007) Learning to integrate arbitrary signals from vision and touch. J Vis 7(5): 7.1-14

Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415(6870):429-33

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

Ernst, M.O. and Bülthoff, H.H. (2004) Merging the senses into a robust percept', Trends Cogn Sci 8(4): 162-9

Fellemann, D.J. and Van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1(1):1-47

Fine, I., Jacobs, R.A. (1999) Modeling the combination of motion, stereo, and vergence angle cues to visual depth. Neural Comput 11(6): 1297-330

Finney, E.M., Fine, I., Dobkins, K.R. (2001) Visual stimuli activate auditory cortex in the deaf. Nat Neurosci. 4(12):1171-3.

Fiser, J. and Aslin, R.N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological Science 12, 499–504

Foxe, J.J., Morocz, I.A., Murray, M.M., Higgins, B.A., Javitt, D.C., Schroeder, C.E. (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10(1-2):77-83

Frens, M.A., Van Opstal, A.J., Van der Willigen, R.F. (1995) Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept Psychophys 57(6):802-16

Gallistel, C.R., Fairhurst, S., Balsam, P. (2004) The learning curve: Implications of a quantitative analysis. Proc Natl Acad Sci USA. 101(36):13124-31

Geisler, W.S., Perry, J.S., Super, B.J., Gallogly, D.P. (2001) Edge co-occurrence in natural images predicts contour grouping performance. Vision Res, 41(6):711-724

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K. (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J Neurosci 25(20):5004-12

Gibson, J.R. and Maunsell, J.H. (1997) Sensory modality specificity of neural activity related to memory in visual cortex. J Neurophysiol 78(3):1263-75

Gielen, S.C., Schmidt, R.A., Van den Heuvel, P.J. (1983) On the nature of intersensory facilitation of reaction time. Percept Psychophys 34(2):161-8

Gold, J.I. and Shadlen, M.N. (2001) Neural computations that underlie decisions about sensory stimuli. Trends Cog Sci 5: 10-16

Gori, M., Del Viva, M., Sandini, G., Burr, D.C. (2008) Young children do not integrate visual and haptic form information. Curr Biol 18(9):694-8

Greenwald, H.S. and Knill, D.C. (2009) A comparison of visuomotor cue integration strategies for object placement and prehension. Vis Neurosci 26(1):63-72

Hagen, M.C., Franzén, O., McGlone, F., Essick, G., Dancer, C., Pardo, J.V. (2002) Tactile motion activates the human middle temporal/V5 (MT/V5) complex. Eur J Neurosci 16(5):957-64

Hairston, W.D., Wallace, M.T., Vaughan, J.W., Stein, B.E., Norris, J.L., Schirillo, J.A. (2003) Visual localization ability influences cross-modal bias. J Cogn Neurosci 15(1):20-9

Helmholtz, H. von (1867) Handbuch der physiologischen Optik. Brockhaus, Leipzig

Hershenson, M. (1962) Reaction time as a measure of intersensory facilitation. J Exp Psychol 63:289-93

Hinton, G. E., Sejnowski, T. J. (1986) Learning and relearning in Boltzmann machines, In: Rumelhart D E and McClelland J L editors Parallel Distributed Processing Explorations in the Microstructure of Cognition Volume Foundations MIT Press Cambridge MA

Hoyer, P. O., Hyvärinen, A. (2003) Interpreting neural response variability as Monte Carlo sampling of the posterior, In Advances in Neural Information Processing Systems 15 (NIPS*2002), pp. 277-284, MIT Press

Jacobs, R.A. (1999) Optimal integration of texture and motion cues to depth. Vision Res 39(21): 3621–3629

Jacobs, R.A. and Fine, I. (1999) Experience-dependent integration of texture and motion cues to depth. Vision Res 39(24): 4062 – 4075

James, T.W., Humphrey, G.K., Gati, J.S., Servos, P., Menon, R.S., Goodale, M.A. (2002) Haptic study of three-dimensional objects activates extrastriate visual areas. Neuropsychologia 40(10): 1706-14

Jones E.G. and Powell T.P. (1970) An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. Brain 93(4): 793-820

Jousmaki, V. and Hari, R. (1998) Parchment-skin illusion: sound-biased touch. Curr Biol 8(6): R190-R191

Kersten, D. (1999) High-level vision as statistical inference. in The New Cognitive Neurosciences, 2nd Edition, Gazzaniga (Ed.), MIT Press, Cambridge

Kahneman, D. and Tversky, A. (2000) Choices, Values, and Frames, Cambridge University Press

Kersten, D., Mamassian, P., Yuille, A. (2004) Object perception as Bayesian

Inference. Annu Rev Psychol 55: 271-304

Kersten, D., Yuille, A. (2003) Bayesian models of object perception. Curr Opin Neurobiol 13(2): 150-8

Knill, D.C. (2003) Mixture models and the probabilistic structure of depth cues. Vision Res 43(7): 831-854

Knill, D.C. (2007) Learning Bayesian priors for depth perception. J Vis 7(8): 13

Knill, D.C. and Kersten, D. (1991) Apparent surface curvature affects lightness perception. Nature 351(6323): 228-30

Knill, D.C. and Saunders, J.A. (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43(24): 2539-58

Knill, D.C. and Pouget, A. (2004) The Bayesian Brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27(12): 712-9

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

Knutsen, P.M. and Ahissar, E. (2008) Orthogonal coding of object location, Trends Neurosci 32(2): 101-109

Koerding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., Shams, L. (2007) Causal inference in multisensory perception. PLoS One 2(9): e943

Körding, K.P., Wolpert, D. (2004) Bayesian Integration in Sensorimotor Learning, Nature 427:244-247

Kujala, T., Huotilainen, M., Sinkkonen, J., Ahonen, A.I., Alho, K., Hämäläinen, M.S., Ilmoniemi, R.J., Kajola, M., Knuutila, J.E., Lavikainen, J., Salonend, O., Simolab, J., Standertskjöld-Nordenstamd, C-G., Tiitinena, H., Tissarie, S.O., Näätänen, R. (1995) Visual cortex activation in blind humans during sound discrimination. Neurosci Lett 183(1-2): 143-6

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 Res 35(3): 389-412

Lewkowicz, D.J. (2000). Perceptual development in human infants. Am. J. Psychol 113(3): 488-499

Lomo, T. and Mollica, A. (1959) Activity of single units of the primary optic cortex during stimulation by light, sound, smell and pain, in unanesthetized rabbits. Boll Soc Ital Biol Sper 35: 1879-82

Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A. (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9(11): 1432-8

MacKay, D. (2003) Information Theory, Inference, and Learning Algorithms.

Cambridge University Press

Mamassian, P., Knill, D.C., Kersten, D. (1998) The perception of cast shadows. Trends Cogn Sci 2(8): 288-295

Mamassian, P. and Landy, M.S. (2001) Interaction of visual prior constraints. Vision Res 41(20): 2653-68

Marr, D. (1982) Vision: a computational investigation into the human representation and processing of visual information. San Francisco, W.H. Freeman & Co.

McGurk, H. and MacDonald, J. (1976) Hearing lips and seeing voices. Nature 264(5588): 746-8

Meredith, M.A., Stein, B.E. (1986) Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res 365(2): 350-4

Michel, M.M., Jacobs, R.A. (2007) Parameter learning but not structure learning: a Bayesian network model of constraints on early perceptual learning. J Vis 7(1): 4

Morrell, F. (1972) Visual System's View of Acoustic Space Nature, 238, 44 - 46

Murata, K., Cramer, H., Bach-y-Rita, P. (1965) Neuronal convergence of noxious, acoustic, and visual stimuli in the visual cortex of the cat. J Neurophysiol 28(6):1223-39

Nardini, M., Jones, P., Bedford, R., Braddick, O. (2006) Development of Cue Integration in Human Navigation. Curr Biol 18(9): 689-93

Neumann J. v., Morgenstern, O. (1944): Theory of games and economic behavior. Princeton: Princeton University Press. 648 p.

Newell, F.N., Ernst, M.O., Tjan, B.S., Bülthoff, H.H. (2001) Viewpoint dependence in visual and haptic object recognition. Psychol Sci 12(1): 37-42

Oruç, I., Maloney, L.T., Landy, M.S. (2003) Weighted linear cue combination with possibly correlated error. Vision Res 43(23): 2451-68

Patton, P.E., Anastasio, T.J. (2003) Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. Neural Comput 15(4): 783-810

Pearl, J. (1988): Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference 2nd Edn. San Mateo: Morgan Kaufmann Publishers

Pick, H.L., Warren, D.H., Hay, J.C. (1969): Sensory conflict in judgements of spatial direction. Percept Psychophys 6: 203-205

Poremba, A., Saunders, R.C., Crane, A.M., Cook, M., Sokoloff, L., Mishkin, M. (2003) Functional mapping of the primate auditory system. Science 299(5606):568-72

Rothkopf, C.A. and Ballard, D.H. (2009) Image statistics at the point of gaze during

human navigation. Vis Neurosci 26(1): 81-92

C. A. Rothkopf, T. H. Weisswange, J. Triesch (2009) Learning independent causes in natural images explains the spacevariant oblique effect, IEEE 8th International Conference on Development and Learning, June 5-7

Rowland, B.A., Stanford, T.R., Stein, B.E. (2007) A model of the neural mechanisms underlying multisensory integration in the superior colliculus. Perception 36(10): 1431-43

Sadato, N., Pascual-Leone, A., Grafman, J., Ibañez, V., Deiber, M.P., Dold, G., Hallett, M. (1996) Activation of the primary visual cortex by Braille reading in blind subjects. Nature 380(6574): 526-8

Sanborn, A., Griffiths, T., Navarro, D. A. (2006) A more rational model of categorization. Proceedings of CogSci 2006: 726-31

Sato, Y., Toyoizumi, T., Aihara, K. (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Comput 19(12): 3335-55

Saunders, J.A. and Knill, D.C. (2001) Perception of 3d surface orientation from skew symmetry. Vision Res 41(24): 3163-3183

Schlicht, E.J., Schrater, P.R. (2007) Effects of visual uncertainty on grasping movements. Exp Brain Res 182(1): 47-57

Schrater, P. R., Kersten, D. (2000) How Optimal Depth Cue Integration Depends on the Task. International Journal of Computer Vision, 40(1): 71-89

Schroeder, C.E. and Foxe, J.J. (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Brain Res Cogn Brain Res 14(1): 187-98

Shams, L. and Seitz, A. R. (2008): Benefits of multisensory learning, Trends in Cognitive Sciences, 12(11): 411-417

Smith, A. M., ed. and trans. (2001), Alhacen's Theory of Visual Perception: A Critical Edition, Transactions of the American Philosophical Society, Philadelphia, 91 (4–5)

Stein, B.E. and Meredith, M.A. (1993) The merging of the senses. Cambridge: MIT Press

Stocker, A. A. and Simoncelli, E. P. (2006) Noise characteristics and prior expectations in human visual speed perception, Nature Neuroscience, vol.9, no.4, p. 578-585

Thomas, G. (1941) Experimental study of the influence of vision on sound localisation. J Exp Psychol 28: 167177

Triesch, J., Ballard, D.H., Jacobs, R.A. (2002) Fast temporal dynamics of visual cue integration. Perception 31(4): 421-34

Triesch, J., von der Malsburg, C. (2001) Democratic integration: self-organized integration of adaptive cues. Neural Comput 13(9): 2049-74

Trommershäuser, J., Maloney, L.T., Landy, M.S. (2003) Statistical decision theory and trade-offs in the control of motor response. Spat Vis 16(3-4): 255-75

Trommershäuser, J., Maloney, L. T., Landy M. S. (2008) Decision making, movement planning and statistical decision theory. Trends in Cognitive Science, 12(8), 291-297

van Beers, R.J., Sittig, A.C., Gon, J.J. (1999) Integration of proprioceptive and visual position-information: an experimentally supported model. J Neurophysiol 81(3): 1355-64

von Schiller, P. (1932) Die Rauhigkeit als intermodale Erscheinung. Z Psychol Bd, 127: 265-289

Wallace, M.T., Meredith, M.A., Stein, B.E. (1992) Integration of multiple sensory modalities in cat cortex. Exp Brain Res 91(3): 484-8

Wallace, M.T., Wilkinson, L.K., Stein, B.E. (1996) Representation and integration of multiple sensory inputs in primate superior colliculus. J Neurophysiol 76(2): 1246-66

Wallace M.T., Stein B.E. (2007) Early experience determines how the senses will interact. Journal of Neurophysiology, 97(1):921-926

Weiss, Y. and Fleet, D.J. (2002) Velocity likelihoods in biological and machine vision. in Probabilistic models of the brain Editors: Rao, Olshausen, Lewicki, MIT Press, Cambridge

Weiss, Y., Simoncelli, E.P., Adelson, E.H. (2002) Motion illusions as optimal percepts. Nat Neurosci 5(6): 598-604

Weisswange, T.H., Rothkopf, C.A., Rodemann, T., Triesch, J. (2009) Can reinforcement learning explain the development of causal inference in multisensory integration? in Proceedings of ICDL 2009

Wozny, D.R., Beierholm, U.R., Shams, L. (2008) Human trimodal perception follows optimal statistical inference. J Vis 8(3): 24, 1-11

Yuille, A.L. and Bülthoff, H.H. (1996): Bayesian Theory and Psychophysics, in Perception as Bayesian Inference, Editors: D. Knill, W. Richards, 123-161, Cambridge University Press

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10(7): 301-308

Zemel, R.S., Dayan, P., Pouget, A. (1998) Probabilistic interpretation of population code. Neural Comput 10(2): 403-30

Zhou, Y.D. and Fuster, J.M. (2000) Visuo-tactile cross-modal associations in cortical somatosensory cells. Proc Natl Acad Sci USA 97(17): 9777-82