Review TRENDS in Cognitive Sciences Vol.8 No.4 April 2004

Merging the senses into a robust percept

Marc O. Ernst and Heinrich H. Bülthoff

Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany

To perceive the external environment our brain uses multiple sources of sensory information derived from several different modalities, including vision, touch and audition. All these different sources of information have to be efficiently merged to form a coherent and robust percept. Here we highlight some of the mechanisms that underlie this merging of the senses in the brain. We show that, depending on the type of information, different combination and integration strategies are used and that prior knowledge is often required for interpreting the sensory signals.

The key to robust perception is the combination and integration of multiple sources of sensory information. This is because no information-processing system, neither technical nor biological, is powerful enough to 'perceive and act' accurately under all conditions. Evidence from a range of psychophysical methods is converging to shed light on how humans achieve robust perception through the combination and integration of information from multiple sensory modalities. This review describes evidence that humans combine information following two general strategies: The first is to maximize information delivered from the different sensory modalities ('sensory combination'). The second strategy is to reduce the variance in the sensory estimate to increase its reliability ('sensory integration') (see also Box 1).

Sensory combination

With seemingly no effort, the human brain reconstructs the environment from the incoming stream of – often ambiguous – sensory information and generates unambiguous interpretations of the world. To do so many different sources of sensory information are constantly processed, analysed and combined. For example, we have all at some time sat on a train looking out the window at a neighbouring train. If the other train starts moving, there is an ambiguous situation: is it your or the other train that is moving? In either case the brain will come up with a unique – right or wrong – answer to this ambiguous situation. If the brain is wrong the illusory self-motion is noticed either when looking out of another window or when a different sensory modality such as the vestibular system disambiguates the situation. That is, the brain collects more and more information about the perceptual event and finally resolves the ambiguity.

Corresponding author: Marc O. Ernst (marc.ernst@tuebingen.mpg.de).

Another illustration of such 'disambiguation' is shown in Figure 1. The bi-stable percept of the Necker Cube (Figure 1a) can easily be disambiguated by adding shadows or a small bar that introduces an occlusion cue (Figure 1b). There are plenty of other examples of disambiguation from within or across the modalities; some of these involve shadows [1], shape from shading [2], specularities [3], or other shape cues. Disambiguation is a way of sensory interaction that our brain uses to form a more robust perceptual estimate [4].

Perception is multisensory [5]; if a single modality is not enough to come up with a robust estimate, information from several modalities can be combined. For example, Newell *et al.* [6] showed that for object recognition different modalities complement each other with the effect of increasing the information content. They showed that both visual and haptic object recognition is dependent on the orientation of the object relative to the observer. The best view for recognizing an object visually is the side that corresponds to the learned view (usually the front). The side that is most accurately recognized by the haptic modality, however, is the side the fingers explore the most. Given natural exploration behaviour of hand-sized objects, this is most often the back. With this natural exploration behaviour the two modalities complement each other ('sensory cooperation' [4]), resulting in an increase of information gathered about the object's shape [7]. This naturally leads to a more robust estimate of the environmental property in question.

Using prior information

No single sensory signal can provide reliable information about the three-dimensional structure of the environment in all circumstances. This incompleteness might be resolved by collecting more and more information using different sources. But rather than delaying an uncertain decision, the brain at any given moment picks a single solution from all the possibilities. A decision is needed to interact with the environment – the main purpose served by the perceptual system. But how does the brain come up with a decision given the ill-posed problem of perception? To resolve ambiguities, the brain uses constraints in the way in which information is used and a knowledge base of previously acquired information (see Box 2). For example, consider the Necker Cube. The 2D pattern of light that the Necker Cube gives rise to on the retina could be caused by an infinite number of 3D structures in the world (Figure 1c). To solve the problem the visual system

 1364-6613/$ - see front matter doi:10.1016/j.tics.2004.02.002

Box 1. What is a cue? Combination versus integration

'Cue combination' is an expression often used to describe interactions between different sources of sensory information. The problem is that there is no precise definition of a 'cue' [51,52]. The cue concept is often dealt with as implicitly understood. For example, most people agree that cues to visual depth include perspective signals, disparity, shadows, shading, motion parallax and occlusion. And cues from touch and audition can also provide depth information (Figure I). However, because of the lack of a precise definition of a 'cue' there is the potential for confusion, especially when talking about cue combination and what rules to apply [4,41,53–55].

We will not claim to be able to provide a clear definition here. However, if we talk about a 'cue' we think of it as any sensory information that gives rise to a sensory estimate. Compared with the common understanding, this view omits the assumption that cues have to be somehow 'independent modules'. For the purpose of this article, to get around the problem of definition of a 'cue' we have divided cue combination into two parts that we refer to as 'sensory combination' and 'sensory integration'.

'Sensory combination' describes interactions between sensory signals that are not redundant. That is, they may be in different units, coordinate systems, or about complementary aspects of the same environmental property. 'Disambiguation' and 'cooperation' are examples for two such interactions. Some of the interactions found within the sensory combination framework are in agreement with strong coupling (fusion) between sensory signals [54,56].

By contrast, 'sensory integration' describes interactions between redundant signals. That is, to be integrated, the sensory estimates must be in the same units, the same coordinates and about the same aspect of the environmental property (Figure I). The Maximum Likelihood Estimate (MLE) model described in the main text is in agreement with the modified weak fusion idea [41].

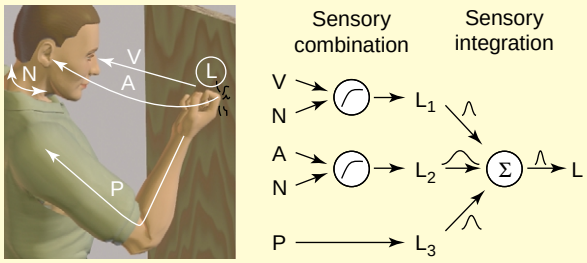

Figure I. When knocking on wood at least three sensory estimates about the location (L) of the knocking event can be derived: visual (V), auditory (A) and proprioceptive (P). In order for these three location signals to be integrated they first have to be transformed into the same coordinates and units. For this, the visual and auditory signals have to be combined with the proprioceptive neck-muscle signals (N) to be transformed into body coordinates (for simplicity we ignore the eye-muscle signals). The process of sensory combination might be non-linear. When sensory combination results in multiple estimates about the same object or event this process is referred to as promotion [41]. At a later stage the three signals (L_1, L_2, L_3) are then integrated to form a coherent percept of the location of the knocking event. Assuming the MLE model for integrating the signals this later stage should be linear.

computes the most likely 3D structure that created the 2D pattern, given previous experience with 3D objects. Shape priors like compactness and regularity can significantly reduce the interpretation space. Two interpretations of the Necker Cube are about equally likely, which is why we alternate between these two interpretations. However, the 'best guess' can also be wrong sometimes and can lead to interesting illusions [8].

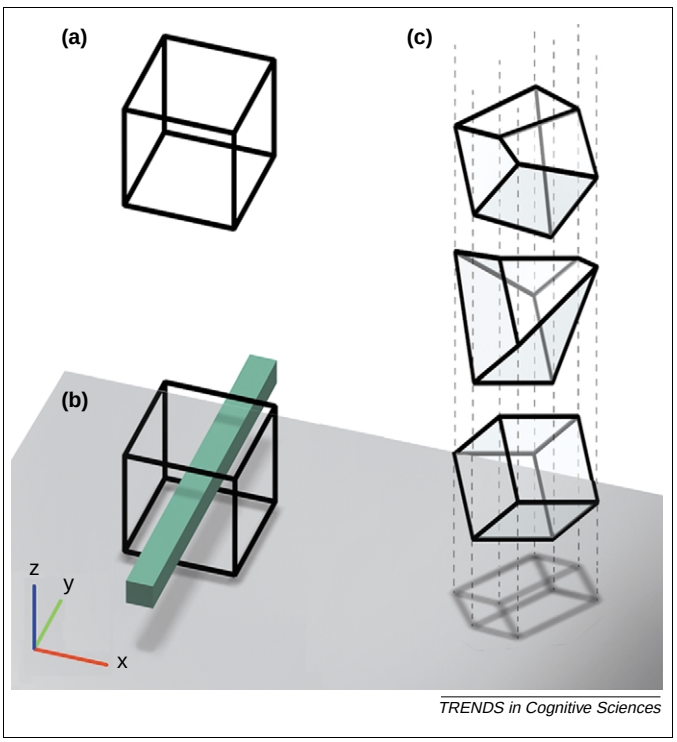

Figure 1. **(a)** The Necker Cube induces a bi-stable percept. **(b)** Disambiguation of the bi-stable Necker Cube percept by introducing an occlusion cue and a shadow. **(c)** An infinite number of 3D configurations could produce the same projection image. Here this fact is illustrated by the cast shadow on the tabletop, but the same projected images would be formed on the eye's retina.

Sensory integration

Often there is more than one sensory estimate available for perceiving some environmental property. For example, when judging an object's size both the visual and haptic modalities can provide information. But what is the perceived size of an object that is simultaneously seen and touched? Is it the one determined by the visual estimate, the one determined by the haptic estimate, or something in-between? Information from the different sensory modalities has to be integrated such that a coherent multisensory percept is formed. What is the mechanism underlying this integration and at what level is it performed?

Visual capture

Rock and Victor conducted a classic experiment in the 1960s that investigated the integration of visual and haptic information [9]. They asked subjects to report the perceived size of an object simultaneously seen and felt. Subjects looked at the object through a cylinder lens that made a square look like a rectangle and that so created a conflict between visual and haptic information. Whether subjects reported perceived size by drawing, visual matching or haptic matching, vision dominated the integrated percept. However, there was also always a small but consistent influence of touch on the integrated percept. This phenomenon of visual dominance was subsequently called 'visual capture'.

Box 2. Perception–action and Bayes' theory

We perceive in order to act and our actions affect the percept of the environment. This action–perception loop is illustrated in Figure I. To allow interactions a reconstruction of the environment based on sensory data has to be formed in the brain. It is however impossible to reconstruct the environment 'bottom-up' from the sensory information alone. Prior knowledge is needed to interpret ambiguous sensory information. Bayesian inference provides a formal way to describe such interactions and enables one to model the uncertainty about the world by combining prior knowledge (that might be unconscious) with observational, sensory evidence (the likelihood function) to infer the most probable interpretation of the environment [57–59].

The Bayesian framework can be used to construct 'ideal observer' models as a standard for comparison with human performance. Bayes' Rule says that the posterior probability $p(W|I)$ is proportional to the product of the likelihood function $p(I|W)$ and prior probability distribution $p(W)$: $p(W|I) \propto p(I|W) \times p(W)$. All the work in the modelling is in specifying the likelihood functions and priors.

The integration of sensory information is mostly described in this article as a bottom-up process. As such it can be modelled using likelihood functions only (thereby ignoring prior distributions or assuming that they are uniform over a wide range, disappearing at infinity). Hence, the integration model discussed is also referred to as Maximum Likelihood Estimation (MLE). We discuss how these models can be extended using prior knowledge. For example, we consider incorporating prior knowledge for disambiguation of sensory information.

In most studies discussed the decision process is assumed to be unbiased and ideal for the selected goal. For a complete analysis of the task in question, however, the decision-making process has to be considered in addition to the sensory-estimation process. Here, the goal for the task is defined using gain and loss functions [60]. Trommershäuser, Maloney and Landy [61] showed that statistical decision theory can be used to accurately explain pointing behaviour for different loss functions. Similarly, Triesch *et al.* [62] showed how the statistical reliability of the stimuli affects the decision process. That is, a complete model has to consider all three parts that make up Bayesian' Decision theory: sensory estimation that includes prior knowledge together with the decision-making process (e.g. [58,63]).

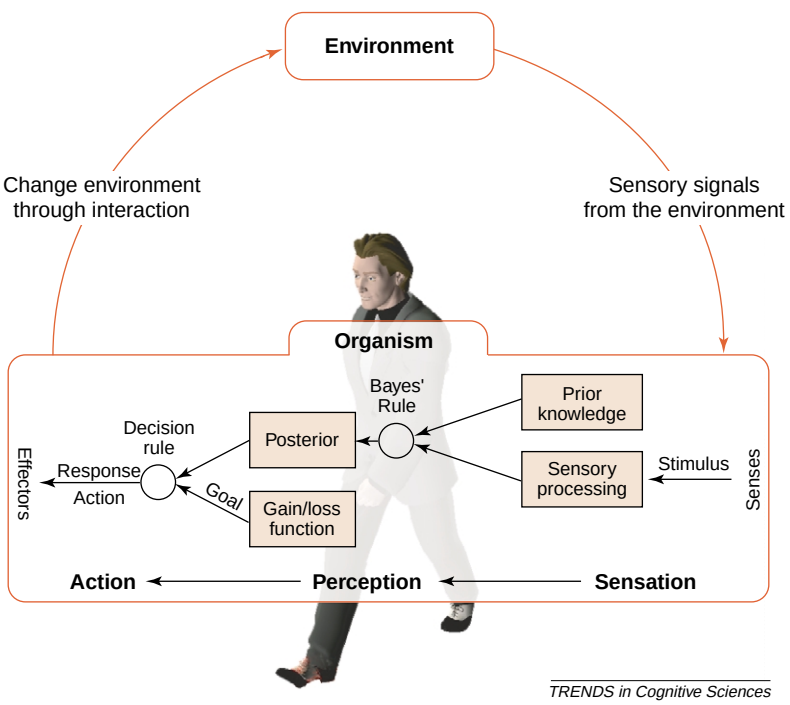

Figure I. The perception–action loop, incorporating a Bayesian framework. (See text for details.)

The visual modality does not always win in such crossmodal tasks. For example, Shams, Kamitani and Shimojo [10] presented subjects with a briefly flashed visual stimulus that was accompanied by one, two or more auditory beeps. There was a clear influence of the number of auditory beeps on the perceived number of visual flashes. That is, if there were two beeps subjects frequently reported seeing two flashes when only one was presented. Maintaining the terminology above, this effect may be called 'auditory capture'.

The 'Modality Precision' or 'Modality Appropriateness' hypothesis by Welch and Warren [11] is often cited when trying to explain which modality dominates under what circumstances. These hypotheses state that discrepancies are always resolved in favour of the more precise or more appropriate modality. In spatial tasks, for example, the visual modality usually dominates, because it is the most precise at determining spatial information. For temporal judgments, however, such as that studied by Shams *et al.* [10] and others (e.g. [12]), the situation is reversed and audition, being the more appropriate modality, usually dominates over vision. However, the terminology used, 'modality precision' and 'modality appropriateness', is misleading because it is not the modality itself or the stimulus that dominates. Rather, the dominance is determined by the estimate and how reliably it can be derived within a specific modality from a given stimulus. Therefore, the term 'estimate precision' would probably be more appropriate.

The Maximum Likelihood Estimation (MLE) model

What would be the most efficient manner to integrate sources of sensory information? First, the goal of sensory estimation must be specified. If the goal is to come up with the most reliable (unbiased) estimate, then the variance of the final estimate should be reduced as much as possible. Every sensory signal is noisy and therefore so is the sensory estimate. That is, if the system made 10 consecutive estimates of exactly the same environmental property, all 10 answers would be slightly different. One cause for this might be the inherent noise in neural transmission (e.g. owing to spontaneous firing).

Given that the noise of individual estimates are independent and Gaussian (assumptions that make the modelling particularly easy), the estimate with the lowest variance is the Maximum Likelihood Estimate (MLE). Thus, the integrated estimate, $\hat{s}$, is the weighted sum of the individual estimates (Eqn 1) with weights w_i proportional to their inverse variances σ_i^2 (Eqn 2) [13]. The index i refers to the different sensory signals. This estimation scheme is also known as a form of the Kalman filter [14,15].

$$\hat{s} = \sum_i w_i \hat{s}_i \text{ with } \sum_i w_i = 1 \qquad \text{(Eqn 1)}$$

$$w_j = \frac{1/\sigma_j^2}{\sum_{i=1\ldots j,\ldots N} 1/\sigma_i^2} \qquad \text{(Eqn 2)}$$

Depending on the situation, the sensory signal's quality can vary and the sensory estimate thus has different measurement reliability (noise). The estimate's weight should take into account the quality of information. If we define the reliability r as the inverse variance of the estimates:

$$r_i = 1/\sigma_i^2, \qquad \text{(Eqn 3)}$$

then the reliability of the integrated estimate given the choice of weights from Eqn 2 is simply the sum of the reliabilities of the individual estimates.

$$r = \sum_i r_i \qquad \text{(Eqn 4)}$$

By integrating sensory information in this 'optimal' way, the reliability of the integrated estimate is increased and yields the most reliable unbiased estimate possible (estimate with minimal variance). Even an unbiased non-linear model cannot reduce the variance more given the constraints stated above [16].

It is interesting to note that even if we did not choose optimal weights we could still benefit from integration. In such a situation the reliability of the integrated estimate would not be maximal. However, it might still be more reliable then the individual estimates [16].

There are several recent studies, reviewed below, showing that humans integrate information both within and across sensory modalities in just such an efficient way. That is, people use the linear weighting rule to integrate signals, where weights depend on the signal's reliability.

Weighting of sensory information

Within the visual modality there are several studies, mostly dealing with depth perception, that confirm that human behaviour is consistent with the linear weighting model (Eqn 1). For example, Johnston, Cumming and Landy [17] found that people form a weighted average of motion and disparity signals when asked to report an object's shape. The same is true for texture and disparity signals to depth [18], for the visual perception of slant [19], for the judgment of texture-defined edges [20], and for the estimate of distance [21]. More importantly, Young *et al.* [18], among others, showed that the weights change in the predicted direction as signal reliability is manipulated. This demonstrates that multiple information sources are used for the associated judgments.

There is also weighting of sensory signals within the haptic modality (e.g. for force and position cues to shape [22,23]) or across the different modalities (e.g. for vision–audition [24], vision–haptic [9] or vision–proprioception [25]). For example van Beers, Sittig and van der Gon [25] investigated visual–proprioceptive integration for localization and found that vision is more precise for discriminations along the horizontal compared with depth discrimination. By contrast, proprioception is more reliable for discriminations along the direction parallel to the forearm (in depth) and worse for the horizontal. The prediction from these asymmetric reliability ellipses is that the MLE should lie on a curved path between the two visually and proprioceptively specified locations and not, as one might naively assume, on a straight path. In their report, van Beers *et al.* provide some evidence for this qualitative prediction.

All the above-mentioned studies considered only the weighting of sensory signals. However, the main purpose of sensory integration is to make the estimates more reliable [26]. That is, there should be an observable reduction in variance compared with the individual estimates (see Eqns 3,4). Moreover, the weighting measure alone does not reveal whether the integration of sensory information is optimal because such data are also consistent with a strategy in which an observer bases an answer on only one cue at a time but switches the response to a cue in proportion to its reliability [20,24]. When averaging the answers this 'cue-switching' strategy mimics the performance of cue weighting. Cue-switching, however, cannot improve performance relative to a single-cue task. On this basis, when investigating reliabilities, it is possible to discriminate cue-switching behaviour from sensory integration.

Reducing variance by integrating signals

Jacobs used a matching task to provide some evidence for variance reduction [27]. This study cannot be taken as definitive, however, because to make predictions for optimal integration behaviour using multiple information sources, the reliability of each individual signal has to be known. The problem in determining these reliabilities is that they are often not independent. In the Jacobs study the stimuli were computer-generated cylinders defined by texture and motion signals. It is impossible to measure directly the reliability for the motion signal in isolation

from texture because there must be some moving texture elements in the stimulus to generate motion. Of course, such texture elements also provide some shape information. Jacobs and others (e.g. [28]) tried to minimize this problem by weakening the texture cue when assessing the 'motion-alone' reliability. However, it is difficult to determine how far this is possible.

Independent sensory signals are needed for investigating optimal MLE-like integration behaviour. The independence assumption is more likely to be true for integration of information across separate modalities than integration within a modality. To test for optimality, in the sense that the integrated unbiased estimate is most reliable, Ernst and Banks [29] investigated visual and haptic discrimination of object size using the set-up illustrated in Figure 2. They determined the reliabilities for discriminating sizes for each modality alone to make predictions for the weights and the integrated reliability in the crossmodal case. To show that weighting changes with the reliability of the signals, Ernst and Banks manipulated the reliability of the visual stimulus by adding noise to the display. Across conditions with different reliabilities the performance in the bimodal task was well predicted by the parameter-free MLE model. The weight changed from visual dominance when there was no noise added to the visual display (so visual information was very reliable) to haptic dominance when there was a lot of added noise (Figure 3). That is, behavior changed smoothly from 'visual capture' to 'haptic capture'. The visual capture here is in agreement with Rock and Victor [9]: under normal conditions with no added noise, vision usually dominates size judgments. Considering the reliability, most benefit should be obtained when the estimates' reliabilities are equal, a fact that was also experimentally confirmed [29]. As all conditions in this study were intermixed, such a reduction in variance is consistent with dynamic trial-by-trial adjustment of weights. This suggests that the nervous system has on-line access to sensory reliabilities.

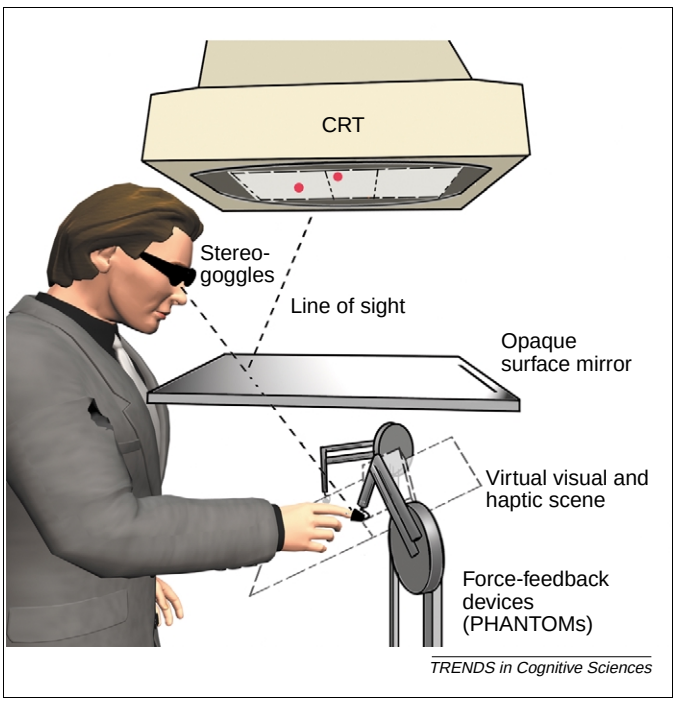

Figure 2. In this visual–haptic set-up used by Ernst and Banks [29] observers view the reflection of the visual stimulus binocularly in a mirror using stereo-goggles. The haptic stimulus is presented with two PHANTOM™ force-feedback devices, one each for the index finger and thumb of the right hand. With this arrangement the visual and the haptic virtual scenes can be independently manipulated.

Ernst and Banks added noise to the display to vary the cues reliabilities. This presents no problem provided that no bias is introduced by this manipulation; all that matters is that adding noise changes the reliability of the signal. Gepshtein and Banks [30] performed a more subtle manipulation to vary the signals' reliability more naturally. They investigated the integration of estimated visual and haptic distance between two semitransparent plates rotated in depth. The haptic estimate's reliability was essentially unchanged with different rotations of the plates in depth. However, the visual estimate was much more reliable when the plates are viewed edge on, in which case the judgment is an estimate of width, than when the plates were viewed through each other, in which case the estimate becomes a judgment in depth. In accordance with the MLE model, the visual modality is dominant in the first but not the second case, where the haptic modality dominates the percept. This study shows that the contextual conditions for stimulus presentation are taken into account when the brain integrates sensory information.

An example of optimal visual and auditory integration for localization is provided by Alais and Burr [31]. Their crossmodal data are in good agreement with the predicted weighting of signals and a reduction of the estimate's variance. It is worth noting that the data for each individual subject were consistent with the prediction of optimal integration. Similar results were found for the integration of stereo and texture signals within vision [32].

Neural models for sensory integration

To integrate signals optimally, the observer has to know the variances of the estimates. There are two possibilities for acquiring this knowledge: either the variances are learned from past experience, or they are determined on-line during the perceptual judgment itself. The first possibility is rather unlikely, considering that there is an infinite number of possible presentation configurations and environmental conditions. Before being able to behave optimally the system would need experience with all these different situations, which requires learning time. Even though generalization or interpolation rules might exist, on-line determination of the estimates' variances seems much more likely.

There are at least two plausible strategies for determining the variance on-line. The variance of a signal might be determined by looking across the fluctuation of responses to a signal, either over some period of time or across a population of independent neurons. Averaging over time is only quasi-online because there has to be a temporal integration window [33]. Therefore, looking across a whole population of neurons would seem to be the most appropriate approach.

As an example of how on-line determination of variances could be achieved using population codes,

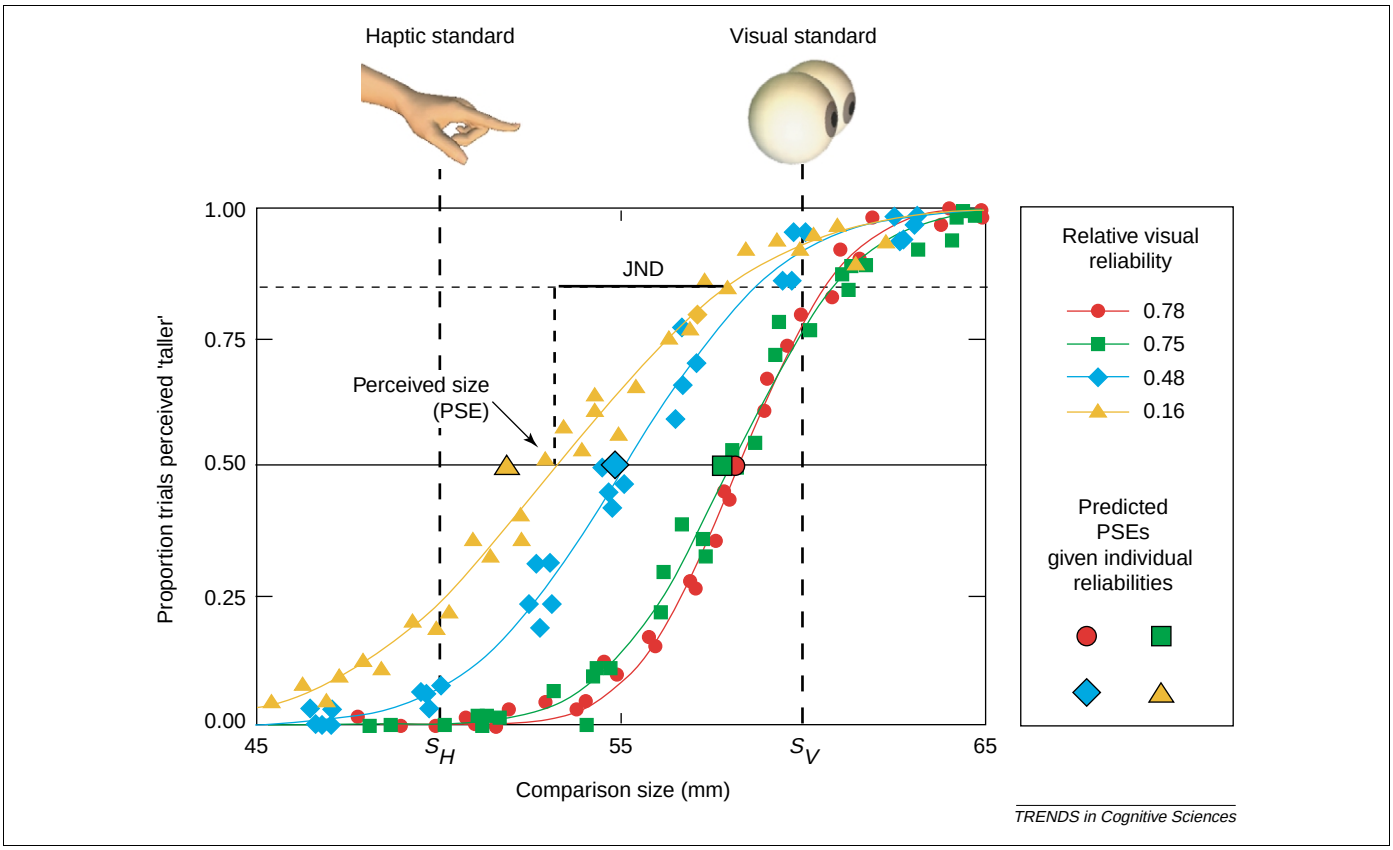

Figure 3. Visual–haptic size-discrimination performance determined with a 2-interval forced-choice task [29]. The relative reliabilities of the individual signals feeding into the combined percept were manipulated by adding noise to the visual display. With these different relative reliabilities four discrimination curves were measured. As the relative visual reliability decreased, the perceived size as indicated by the point of subjective equality (PSE) was increasingly determined by the haptic size estimate (haptic standard, S_H) and less by the visual size estimate (visual standard, S_V). This demonstrates the weighting behaviour the brain adopts and the smooth change from visual dominance (red circles) to haptic dominance (orange triangles). As shown, the PSEs predicted from the individual visual and haptic discrimination performance (larger symbols with black outline) correspond closely to the empirically determined PSEs in the combined visual–haptic discrimination task. (JND = just noticeable difference.) Reproduced from [29], with permission of Nature Publishing Group.

consider a population of neurons sensitive to some object property, say its orientation. Each neuron has a different preferred orientation to which it is maximally tuned and responds less strongly to all other orientations. When stimulated with some orientation, the population activity will have a clearly defined peak marking that orientation and also some variance. For each relevant sensory signal there might be such a population of neurons, and multiplying two such population activities for two signals will result in an overall response that has the characteristics of the MLE integration model [29]. There are several recent developments trying to build realistic neural models for integrating information using population codes (e.g. [29,34–38]).

Beyond sensory integration

The benefit of integrating sources of sensory information is a reduction in the variance of estimates and hence a more reliable percept. However, there must be limits for optimality and conditions under which sensory integration is not the best strategy. In the following we review a few such limits and conditions.

Estimates with correlated noise distributions

One assumption often made when investigating sensory integration is that the noise distributions of the sensory estimates are independent. However, this assumption might not hold in many cases. Landy and Kojima [20], for example, investigated integration of two textural features for the localization of texture-defined edges. Many aspects of their data can be well described by the integration model. However, whether they actually observed optimal performance was somewhat ambiguous. When testing for optimality, Landy and Kojima assumed that the two edge estimates derived from the textural features could be treated as variables with uncorrelated noises. However, this is unlikely to be true because the sensory signals are probably largely processed by the same set of neurons, so that the neuronal noise should be at least partially the same for the two estimates [16]. Other noise sources might also have contributed to the correlation. It seems clear, however, that the potential for correlated noise is higher for the integration of signals within the same modality in comparison with crossmodal integration.

Oruç *et al.* [16] recently investigated the effect that correlated noise distributions have on sensory integration. For the most part, correlated noise will reduce the reliability of integrated estimates and will somewhat alter the weights. However, even with correlated noise, there is generally still a benefit from integrating information.

Discrepant signals and the correspondence problem

It is not always reasonable to integrate sensory signals. For the signals to be integrated the brain has to know

exactly which signals are derived from the same object or event. This is a form of correspondence problem (found also in the motion or stereo literature (e.g. [39,40]), which has to be solved before the signals can be integrated into a uniform percept. We do not know how such a correspondence between signals is established. Signals are most likely to be integrated if they occur simultaneously with no spatial discrepancy, and are not likely to be integrated if the spatial discrepancy is large or if the temporal sequence of events is not appropriate. That is, with large discrepancies robust behavior [41] might be observed in which a discrepant source is discounted or 'vetoed' [4,42] instead of being integrated. But what determines the integration limits?

To investigate optimal integration behaviour, Ernst and Banks [29] among others used a forced-choice discrimination paradigm and asked subjects to compare perceived sizes. Note that subjects are forced to report one number even if they perceive a conflict. If the decision process is optimal (Box 2) and takes the variance of the signals into account, performance would still be optimal and so it is impossible to determine the integration limits with such discrimination tasks. To get around this problem, Hillis *et al.* [43] used an oddity task to investigate how easily changes in the stimulus can be observed. By manipulating the individual signals independently, they found that it is harder to detect discrepancies for combinations of stereo and texture cues to surface slant, than it is for visual and haptic size cues. It is interesting to note that to solve the correspondence problem and to resolve conflicts via adaptation or recalibration, the system at some level has to retain access to the individual estimates. There are several examples in the literature of visual–haptic recalibration (e.g. [44–48]).

Top-down influences and prior assumptions

The MLE approach described above is entirely bottom-up. To incorporate top-down influences or prior assumptions, the model has to be extended. This extension is rather simple if the prior probability can be represented by an independent Gaussian distribution. In this case the prior is just another additional factor in the linear sum of the MLE model (Eqn 1), with a mean that corresponds to the peak of the Gaussian distribution and a weight that is inversely proportional to the squared width of this distribution [16].

Earlier, we demonstrated the use of prior knowledge for sensory combination – but when would using prior knowledge for integrating signals be sensible? The MLE model already accounts for the reliability (i.e. the noise) of the sensory data. However, in addition to having error from noise, sensory estimates can also be biased. This bias may be unstable, for example, because of fast adaptation processes that constantly react to small discrepancies. Such a bias uncertainty would not be reflected in the estimate's noise distribution and therefore would not directly affect the estimate's reliability. However, the brain could learn this bias uncertainty and use this knowledge to emphasize the more stable estimates.

Data from Ernst, Banks and Bülthoff [44] could be interpreted in this way. In their experiment, they investigated the integration of two visual cues to slant – texture and disparity. During a training phase, they introduced tactile feedback that was consistent with one of two visual signals; the other signal was randomly varied. After training, they found a change in perceived slant indicating that the haptically reinforced visual signal was more dominant. Other studies confirm these results [49,50]. From these studies it is impossible to distinguish whether the altered percept was actually due to a change in prior or weight because learning could be manifested either way.

Box 3. Questions for future research

- Can rules be defined for sensory combination as compared with sensory integration?
- Where are the limits on optimal sensory integration behaviour?
- What are the temporal aspects of sensory integration?
- How do top-down influences such as learning, memory and attention affect sensory integration?
- How do we solve the 'correspondence problem' for sensory integration?

Conclusion

In summary, good progress has been made in the past decade in understanding how the brain combines and integrates different sources of information by an interdisciplinary effort of researchers from neuroscience, cognitive science, computer vision, robotics and mathematics (see also Box 3).

We have discussed how the brain reduces the variance in the integrated estimate and increases the robustness of the percept by combining and integrating sources of sensory information from within and across modalities. We argue that to perform optimally the system has to have on-line access not only to the sensory estimate but also to its reliability.

Acknowledgements

This work was supported by the Max Planck Society and by the 5th Framework IST Program of the EU (IST-2001–38040, TOUCH-HapSys). We thank Marty Banks, Roberta Klatzky, Andrew Welchman and Knut Drewing for helpful comments on this draft and Martin Breidt for help with the figures.

References

1 Kersten, D. *et al.* (1996) Illusory motion from shadows. *Nature* 379, 31
2 Bülthoff, H.H. (1991) Shape from X: psychophysics and computation. In *Computational Models of Visual Processing* (Landy, M. and Movshon, A., eds), pp. 305–330, MIT Press
3 Blake, A. and Bülthoff, H.H. (1990) Does the brain know the physics of specular reflection? *Nature* 343, 165–168
4 Bülthoff, H.H. and Mallot, H.A. (1988) Integration of depth modules: stereo and shading. *J. Opt. Soc. Am. A* 5, 1749–1758
5 Stein, B.E. and Meredith, M.A. (1993) *The Merging of the Senses*, MIT Press
6 Newell, F.N. *et al.* (2001) Viewpoint dependence in visual and haptic object recognition. *Psychol. Sci.* 12, 37–42
7 Newell, F.N. *et al.* (2003) Cross-modal perception of actively explored objects. In *Eurohaptics 2003 Conference Proceedings* (O'Modhrain, S. *et al.*, eds), pp. 291–299
8 Sinha, P. and Poggio, T. (1996) The role of learning in 3-D form perception. *Nature* 384, 460–463
9 Rock, I. and Victor, J. (1964) Vision and touch: an experimentally created conflict between the two senses. *Science* 143, 594–596
10 Shams, L. *et al.* (2000) What you see is what you hear. *Nature* 408, 788
11 Welch, R.B. and Warren, D.H. (1986) Intersensory interactions. In

Handbook of Perception and Human Performance (Boff, K.R., *et al.*, eds.), pp. 25.1–25.36, Wiley
12 Spence, C. and Squire, S. (2003) Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, 519–521
13 Cochran, W.G. (1937) Problems arising in the analysis of a series of similar experiments. *J. R. Stat. Soc.* 4, 102–118
14 Kalman, R.E. and Bucy, R.S. (1961) New results in linear filtering and prediction problems. *J. Basic Eng. Ser. D* 83, 95–108
15 Goodwin, G.C. and Sin, K.S. (1984) *Adaptive Filtering Prediction and Control*, Prentice Hall
16 Oruç, I. *et al.* (2003) Weighted linear cue combination with possibly correlated error. *Vis. Res.* 43, 2451–2468
17 Johnston, E.B. *et al.* (1994) Integration of stereopsis and motion shape cues. *Vis. Res.* 34, 2259–2275
18 Young, M.J. *et al.* (1993) A perturbation analysis of depth perception from combinations of texture and motion cues. *Vis. Res.* 33, 2685–2696
19 Backus, B.T. *et al.* (1999) Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vis. Res.* 39, 1143–1170
20 Landy, M.S. and Kojima, H. (2001) Ideal cue combination for localizing texture defined edges. *J. Opt. Soc. Am. A Opt Image Sci. Vis.* 18, 2307–2320
21 Brenner, E. and van Damme, W.J.M. (1999) Perceived distance, shape and size. *Vis. Res.* 39, 975–986
22 Robles-De-La-Torre, G. and Hayward, V. (2001) Force can overcome object geometry in the perception of shape through active touch. *Nature* 412, 445–448
23 Drewing, K. and Ernst, M.O. (2003) Cue integration in the haptic perception of virtual shapes. In *Presence 2003 Conference Proceedings* (Granum, E. *et al.*, eds), p. 39
24 Gharamani, Z. *et al.* (1997) In *Self-organization, Computational Maps, and Motor Control* (Morasso, P.G. and Sanguineti, V., eds), pp. 117–147, Elsevier
25 van Beers, R.J. *et al.* (1999) Integration of proprioceptive and visual position information: an experimentally supported model. *J. Neurophysiol.* 81, 1355–1364
26 Jacobs, R.A. (2002) What determines visual cue reliability. *Trends Cogn. Sci.* 6, 345–350
27 Jacobs, R.A. (1999) Optimal integration of texture and motion cues to depth. *Vis. Res.* 39, 3621–3629
28 Perotti, V.J. *et al.* (1998) The perception of surface curvature from optical motion. *Percept. Psychophys.* 60, 377–388
29 Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433
30 Gepshtein, S. and Banks, M.S. (2003) Viewing geometry determines how vision and haptics combine in size perception. *Curr. Biol.* 13, 483–488
31 Alais, D. and Burr, D. (2004) The ventriloquist effect results from near optimal crossmodal integration. *Curr. Biol.* 14, 257–262
32 Knill, D.C. and Saunders, J.A. (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vis. Res.* 43, 2539–2558
33 Anastasio, T.J. *et al.* (2000) Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* 12, 1165–1187
34 Zemel, R.S. and Dayan, P. (1997) Combining probabilistic population codes. *Int. J. Comput. Artif. Intell.* 15, 1114–1119
35 Deneve, S. *et al.* (1999) Reading population codes: a neural implementation of ideal observers. *Nat. Neurosci.* 2, 740–745
36 Pouget, A. *et al.* (2000) Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132
37 Barber, M.J. *et al.* (2003) Neural representation of probabilistic information. *Neural Comput.* 15, 1843–1864
38 Pouget, A. *et al.* (2003) Computation and inference with population codes. *Annu. Rev. Neurosci.* 26, 381–410
39 Adelson, E.H. and Bergen, J.R. (1985) Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2, 284–299
40 Ohzawa, I. *et al.* (1990) Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science* 249, 1037–1041
41 Landy, M.S. *et al.* (1995) Measurement and modeling of depth cue combination: in defense of weak fusion. *Vis. Res.* 35, 389–412
42 Banks, M.S. and Backus, B.T. (1998) Extra-retinal and perspective cues cause the small range of the induced effect. *Vis. Res.* 38, 187–194
43 Hillis, J.M. *et al.* (2002) Combining sensory information: mandatory fusion within, but not between, senses. *Science* 298, 1627–1630
44 Ernst, M.O. *et al.* (2000) Touch can change visual slant perception. *Nat. Neurosci.* 3, 69–73
45 Epstein, W. and Morgan, C.L. (1970) Adaptation to uniocular image magnification: modification of the disparity-depth relationship. *Am. J. Psychol.* 83, 322–329
46 Adams, W.J. *et al.* (2001) Adaptation to three-dimensional distortions in human vision. *Nat. Neurosci.* 4, 1063–1064
47 De Gelder, B. and Bertelson, P. (2003) Multisensory integration, perception and ecological validity. *Trends Cogn. Sci.* 7, 460–467
48 Atkins, J.E. *et al.* (2003) Experience-dependent visual cue recalibration based on discrepancies between visual and haptic percepts. *Vis. Res.* 43, 2603–2613
49 Atkins, J.E. *et al.* (2001) Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vis. Res.* 41, 449–461
50 Jacobs, R.A. and Fine, I. (1999) Experience-dependent integration of texture and motion cues to depth. *Vis. Res.* 39, 4062–4075
51 Ittleson, W.H. (1960) *Visual Space Perception*, Springer
52 Knill, D.C. *et al.* (1996) Implications of a Bayesian formulation of visual information for processing for psychophysics. In *Perception as Bayesian Inference* (Knill, D.C. and Richards, W., eds), pp. 239–286, Cambridge University Press
53 Massaro, D.W. and Friedman, D. (1990) Models of integration given multiple sources of information. *Psychol. Rev.* 97, 225–252
54 Clark, J.J. and Yuille, A.L. (1990) *Data Fusion for Sensory Information Processing Systems*, Kluwer
55 Howard, I.P. and Rogers, B.J. (2002) Interactions between depth cues. In *Seeing in Depth* (Vol. 2) (Howard, I.P. and Rogers, B.J., eds), pp. 469–493, Porteous
56 Yuille, A.L. and Bülthoff, H.H. (1996) Bayesian theory and psychophysics. In *Perception as Bayesian Inference* (Knill, D. and Richards, W., eds), pp. 123–161, Cambridge University Press
57 Knill, D. and Richards, W. (1996) *Perception as Bayesian Inference*, Cambridge University Press
58 Mamassian, P. *et al.* (2002) Bayesian modelling of visual perception. In *Probabilistic Models of the Brain* (Rao, P.N. *et al.*, eds), pp. 13–36, MIT Press
59 Kersten, D. and Yuille, A.L. (2003) Bayesian models of object perception. *Curr. Opin. Neurobiol.* 13, 150–158
60 Schrater, P.R. and Kersten, D. (2000) How optimal depth cue integration depends on the task. *Int. J. Comput. Vision* 40, 71–89
61 Trommershäuser, J. *et al.* (2003) Statistical decision theory and the selection of rapid, goal-directed movements. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 20, 1419–1433
62 Triesch, J. *et al.* (2002) Fast temporal dynamics of visual cue integration. *Perception* 31, 421–434
63 Kersten, D. (1999) High-level vision as statistical inference. In *The New Cognitive Neurosciences* (Gazzaniga, M.S., ed.), pp. 353–363, MIT Press