Title

Self-Motion Leads to Mandatory Cue Fusion across Sensory Modalities

Authors

Mario Prsa\textsuperscript{1,2}, Steven Gale\textsuperscript{1,2} and Olaf Blanke\textsuperscript{1,2,3}

Author Contributions

M.P. and S.G. contributed to designing and conducting the experiments. M.P. designed and performed the analyses. O.B. supervised the project. M.P. and O.B. wrote the paper.

Affiliation

\textsuperscript{1}Laboratory of Cognitive Neuroscience, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, 1015 Lausanne, Switzerland
\textsuperscript{2}Center for Neuroprosthetics, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
\textsuperscript{3}Department of Neurology, University Hospital Geneva, 1211 Geneva, Switzerland

Running Head

Cross-modal mandatory cue fusion

Contact Information

Mario Prsa
Laboratory of Cognitive Neuroscience
Brain-Mind Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Station 19, Al 2101
1015 Lausanne, Switzerland
mario.prsa@mail.mcgill.ca
Abstract

When perceiving properties of the world we effortlessly combine multiple sensory cues into optimal estimates. Estimates derived from the individual cues are generally retained once the multisensory estimate is produced, and discarded only if the cues stem from the same sensory modality (i.e. mandatory fusion). Does multisensory integration differ in that respect when the object of perception is one's own body rather than an external variable? We quantified how humans combine visual and vestibular information for perceiving own-body rotations and specifically tested whether such idiothetic cues are subjected to mandatory fusion. Participants made extensive size comparisons between successive whole body rotations using only visual, vestibular and both senses together. Probabilistic descriptions of the subjects’ perceptual estimates were compared to a Bayes-optimal integration model. Similarity between model predictions and experimental data echoed a statistically optimal mechanism of multisensory integration. Most importantly, size discrimination data for rotations composed of both stimuli was best accounted for by a model in which only the bimodal estimator is accessible for perceptual judgments as opposed to an independent or additive use of all three estimators (visual, vestibular and bimodal). Indeed, subjects’ thresholds for detecting two multisensory rotations as different from one another were, in pertinent cases, larger than those measured from using either single cue estimate alone. Rotations different in terms of the individual visual and vestibular inputs but quasi-identical in terms of the integrated bimodal estimate became perceptual metamers. This reveals an exceptional case of mandatory fusion of cues stemming from two different sensory modalities.
Introduction

Sensory cues either within (Hillis et al. 2004; Knill and Saunders 2003) or across
(Alais and Burr 2004; Butler et al. 2010; Ernst and Banks 2002; Fetsch et al. 2009;
Gu et al. 2008; Mendonca et al. 2011; van Beers et al. 1999; vanBeers et al. 1996)
senses are often combined such as to produce the final percept according to
statistical optimality. Up until recently, this framework has been exclusively tested in
situations where the object of perception is external to the body of the observer.
However, the observer’s body is also a multisensory object subjected to perceptual
processes (Ionta et al. 2011). This is in particular apparent in passive whole body
displacements that are perceived using mainly vision and the vestibular organs
(Butler et al. 2010; Buttner and Henn 1981; Fetsch et al. 2009; Gu et al. 2008; Young
et al. 1973). The associated cues are said to be idiothetic in nature because they are
derived from the observer’s own displacements. Does the brain process and
integrate two idiothetic signals differently from two externally generated signals? For
example, an observer can simultaneously and independently use vision and audition
to estimate the position of two different objects. In other words, sensory cues making
the same physical measurement are often attributed to different causes and are not
integrated (Koerding et al. 2007; Parise et al. 2012; Shams and Beierholm 2010). It
seems that such dissociations cannot be made in the case of idiothetic cues in most
ecological conditions. Visual and vestibular information relevant for perceiving self-
motion are necessarily redundant and therefore always integrated. Optic flow
incongruent with vestibular input does indeed arise, but is in such cases caused by
constituents of the visual surrounding that are not world-stationary. These visual cues
are not idiothetic, they are attributed to external objects and thus most likely vetoed
when estimating self-motion. Even in the presence of eye movements, optic flow
information corresponding to the actual motion of the body is extracted to guarantee perceptual stability (Haarmeier et al. 2001; Haarmeier et al. 1997; Royden et al. 1992). Because, for the purpose of estimating self-motion, visuo-vestibular integration involves sensory cues providing two ecologically non-dissociable signals, its neural underpinnings must in some respect differ from those underlying non-idiothetic cue integration.

Extending recent work in heading perception involving linear translation stimuli (Butler et al. 2010; Fetsch et al. 2009; Gu et al. 2008) as well as theoretical models of the optimal use of vestibular signals in general (Laurens and Droulez 2007; MacNeilage et al. 2007; Zupan et al. 2002), here we first show that the statistically optimal model of multisensory integration also applies to visual and vestibular cues when perceiving passive self-rotation. Crucially, we subsequently test whether the associated cues are subjected to mandatory fusion. Mandatory fusion entails that, once integrated to produce a more reliable bi-sensory percept, perceptual access to uni-sensory estimates is lost, and has been claimed not to occur across different sensory modalities (Hillis et al. 2002). Here, we demonstrate however that cross-modal mandatory fusion can ensue from idiothetic sensory input. When that situation arises, different cue combinations can theoretically give rise to the same fused percept since they would differ only in terms of information that is lost. For example, a perceived rotation size $S$ borne out by a whole body rotation of size $S + \Delta$ paired with an equally reliable visual cue simulating a rotation of size $S - \Delta$ can be indistinguishable from a true rotation size $S$ produced by both stimuli. Such physically different but perceptually indistinguishable stimuli have been called metamers (Richards 1979) and can be compared to lights of the same color but of
different spectral composition. We argue that cues providing two ecologically non-dissociable signals about own-body displacements account for this phenomenon.

Materials and Methods

The optimal observer model

Based on how probable perceiving an own-body rotation of a certain size is given the visual and vestibular stimuli individually, the multisensory estimate can be predicted from probability theory. To this end, we describe each perceptual estimate in the form of a probability distribution (likelihood). The likelihood provides a probabilistic measure of the estimate; its most likely value and the uncertainty associated with this value. Bayesian statistics then formalize the optimal strategy for combining likelihoods arising from multiple sensory cues (and prior beliefs) to form the a posteriori estimate; the end result of the perceptual process. We describe in this section an optimal observer model, show how it predicts both the variance and the mean of the posterior when human subjects integrate visual and vestibular cues for perceiving whole body rotations, and finally describe within the same framework how to test whether a mandatory fusion of the two cues occurs.

When the body is passively rotated around its yaw axis, both the visual and vestibular senses provide independent information ($l_{vi}$ and $l_{ve}$) about the rotatory stimulus $S$. Probabilistic descriptions of the subjects’ ensuing perception of rotation $\hat{S}$ can be derived from visual and vestibular cues only (i.e. the likelihoods $P(l_{vi}|S)$ and $P(l_{ve}|S)$) and the bimodal pairing (i.e. the posterior $P(S|l_{vi},l_{ve})$). Each distribution recounts how likely it is to perceive any given rotation size. The value corresponding
to the peak of the distribution is the most likely estimate and its standard deviation captures how uncertain/reliable the estimate is (small values indicate low uncertainty and high reliability) (Fig. 1).

Maximum likelihood estimation (MLE), derived from Bayes’ rule (Landy et al. 1995; Yuille and Bulthoff 1996), predicts that the optimal way for the brain to combine sensory cues will result in the posterior distribution being a normalized product of the visual and vestibular likelihoods

$$P(S|I_{vi}, I_{ve}) \propto P(I_{vi}|S)P(I_{ve}|S)$$

(1)

given the assumed Gaussian distributions are produced by independent sources of visual and vestibular noise. The predicted variance of $\hat{S}_p$, the value of $S$ that maximizes $P(S|I_{vi}, I_{ve})$, is then

$$\sigma_p^2 = \frac{\sigma_{vi}^2 \sigma_{ve}^2}{\sigma_{vi}^2 + \sigma_{ve}^2}$$

(2)

where $\sigma_{vi}^2$ and $\sigma_{ve}^2$ are the variances of rotation estimates that maximize $P(S|I_{vi})$ and $P(S|I_{ve})$, respectively. $P(S|I_{vi})$ and $P(S|I_{ve})$ are the single cue posteriors and are equal to the respective likelihoods under the assumption that all sensory signals are equally likely to occur. The maximum a posteriori estimate (value corresponding to the peak of the posterior) will be a weighted average of the two most likely single cue estimates

$$\hat{S}_p = w_{vi}\hat{S}_{vi} + w_{ve}\hat{S}_{ve}$$

(3)

where the weights are equal to the normalized reliabilities (inverse of variance)

$$w_{vi} = \frac{1}{1/\sigma_{vi}^2 + 1/\sigma_{ve}^2} \quad w_{ve} = \frac{1}{1/\sigma_{vi}^2 + 1/\sigma_{ve}^2}$$

(4)
It follows that optimal integration always reduces the variability of the posterior relative to the individual likelihoods (Fig. 1a), and the estimate of the more reliable likelihood weighs more heavily on the posterior estimate (Fig. 1b).

To discriminate between a standard rotation of size $S$ and a test rotation of size $S + \Delta$, the probabilistic observer can compare their visual and vestibular estimates $\hat{p}_{vi}$ and $\hat{p}_{ve}$ on any given trial. We refer to the two as the single cue estimators and we define the single cue discrimination thresholds ($T_{vi}$ and $T_{ve}$) as equal to one standard deviation over trials of $\hat{p}_{vi}$ and $\hat{p}_{ve}$ (i.e. $\sigma_{vi}$ and $\sigma_{ve}$). When the rotations are composed of both stimuli, sensory integration produces an additional bi-sensory estimator $\hat{p}_{bi}$ with discrimination threshold $T_{bi}$ estimated as in Equation 2. Mandatory fusion can be tested by assessing which of the three estimators the subjects use to perceive the bimodal pairs ($S_{vi} = S$, $S_{ve} = S$) and ($S_{vi} = S + \Delta_{vi}$, $S_{ve} = S + \Delta_{ve}$) as different (Fig. 1c-d). If all three estimators are available for perceptual judgments, then the two rotations can be perceived as different if either of the following inequalities is satisfied:

\begin{align}
|\Delta_{vi}| &> T_{vi} \quad (5) \\
|\Delta_{ve}| &> T_{ve} \quad (6) \\
|w_{vi}\Delta_{vi} + w_{ve}\Delta_{ve}| &> T_{bi} \quad (7)
\end{align}

where the cue weights are estimated as in Equation 4. The left hand side of Equation 7 corresponds to the size difference $\Delta_{bi}$ between the bi-sensory estimates of the two rotations. Therefore, if all three estimators are accessible, successful discrimination can be achieved whenever the estimated difference between the two rotations using any of the estimators reaches its own discrimination threshold $\pm T_{vi}$, $\pm T_{ve}$ or $\pm T_{bi}$. 
When the visual and vestibular stimuli rotate by the same amount ($\Delta v_i = \Delta v_e$), an optimal observer can achieve better discrimination performance if using the bi-sensory estimator (Fig. 1c). However, with different combinations of $\Delta v_i$ and $\Delta v_e$ the two rotations can be close to metamers (e.g. if $\Delta v_i = -\Delta v_e$, Fig. 1d), indistinguishable in terms of $\tilde{p}_{bi}$ (the estimated difference using $\hat{p}_{bi}$ will not reach $\pm T_{bi}$), and successful discrimination can be achieved only if access to the unimodal estimators is still available. Empirical assessments of the subjects’ discrimination thresholds and a comparison with theoretical predictions for different $\Delta v_i/\Delta v_e$ values can therefore reveal which estimators the subjects have access to and use for perceiving the two rotations as different.

**Experimental setup**

Vestibular stimuli were delivered in complete darkness by a centrifuge cockpit-style chair digitally servo-controlled (PCI-7352) with highly precise positioning ($\pm 0.1^\circ$). The chair was centered on the rotation axis so that only angular and no linear stimuli were provided to the vestibular organs. Subjects were comfortably restrained with a 5 point racing harness, feet straps and additional cushioning. Head movements were minimized using a head pillow and face paddles pressed against the cheek bones. Rotation profiles were pre-computed and specified the chair’s instantaneous angular position at a rate of 100Hz. The rotations’ velocity profile $v(t)$ was a single cycle of a 0.77Hz raised cosine function (Fig. 2b)

$$v(t) = \frac{A}{T} \left[ 1 - \cos \left( \frac{2\pi t}{T} \right) \right]$$  \hspace{1cm} (8)

where $A$ is rotation size and $T$ its duration ($T = 1.3s$ in this case). Instantaneous angular position $p(t)$ is then specified as
\[ p(t) = A \left[ t - \frac{1}{2\pi} \sin \left( \frac{2\pi t}{T} \right) \right] \]  

(9)

Such transient stimuli fall into the passband range of the transfer function describing the cascade of semi-circular canal dynamics and the velocity storage mechanism (Bertolini et al. 2011). The signal encoding the rotation at the first stage of neural processing is therefore quasi-identical to actual head velocity in this case. Visual only stimuli were simulated by a moving stereoscopic random dot pattern presented on a 22” display fixed to the chair and facing the subject at a distance of approximately 29 cm (Fig. 2a). The limited visual field therefore covered approximately 80° of horizontal and 56° of vertical visual angle. The subject and the display were physically enclosed so as to eliminate any visual cues that might emanate from the stationary surroundings during rotations. The visual scene was constructed as an almost infinite three dimensional volume of randomly distributed dots of different size. Rotations were then simulated by having the observer’s point of view placed in the middle of this space and rotated around the yaw axis. The generated scene simulated therefore retinal optic flow information that would ensue from actual rotation. The stereoscopic stimulus was generated by the Nvidia Quadro FX 3800 graphics card using the OpenGL quad-buffer mechanism. The stimulus was programmed with the Python language and viewed with the Nvidia 3D Vision Kit (active shutter glasses) paired with a Samsung Syncmaster 2233RZ display (120Hz refresh rate) via an infrared transmitter. Subjects were required to maintain visual fixation on a stationary target in the middle of the display in all conditions and masking white noise was delivered over ear phones at all times. The fixation dot was of different color than the random dot pattern and appeared with zero binocular disparity.
Participants

8 healthy adults (M.P., S.G. and 6 naïve to the aims of the experiment) with normal or corrected vision and no history of inner ear disease participated in each experiment (optimal integration: 1 female, mean age 27±5.8; mandatory fusion: 4 female, mean age 24±4.2). In both experiments, an additional subject completed the task but had to be discarded because the performance did not surpass chance level in a number of conditions. All participants gave informed consent and received monetary retribution at 20 CHF/h. The studies were approved by a local ethics committee and were conducted in accordance with the declaration of Helsinki.

Experimental Paradigms

To test for optimal integration, subjects judged the relative size of two successive rotations (the standard and the test) in a two-alternative forced choice task (Fig. 2c). The size of the standard was 15° and the test any of seven equally spaced angles in the interval 10°-20° tested using the method of constant stimuli. These values were chosen based on preliminary tests conducted on two subjects and such that they would include one point on each end of the psychometric fitting curve where size discrimination could be achieved with nearly 100% certainty. The two rotations were preceded, followed and separated by an interval of 0.5s. A 2s period followed during which the subject had to answer, via a button press, whether the second rotation was bigger or smaller than the first. The standard rotation was randomly assigned to either come first or second, but measured responses were always those comparing the test to the standard. Each comparison was repeated 10 times per subject and no feedback was given. Trials in which the subjects failed to give an answer were discarded; this constituted 0.2% of all trials. The relative reliability of the visual and
vestibular cues was manipulated by changing the coherence of the visual motion (number of dots simulating rotation / number of dots moving randomly) from trial to trial between four different levels. The random motion varied in both direction and speed, but the stereoscopic depth of the individual dots did not change during the motion. The four coherence levels used were 100%, 75% (for 2 subjects) or 65% (for 6 subjects), 50% and 25%. For bimodal comparisons, visual and vestibular stimuli were temporally synchronized and occurred simultaneously in congruent (i.e. opposite) directions. Three conflict angles were tested in the bimodal case ($S_{ve}$-$S_{ve}$=0°, 4° or -4°) and were applied to the standard rotation. This produced 17 conditions in total (1 vestibular, 4 visual and 12 bimodal) giving rise to 1190 trials per subject (7 angles x 10 repetitions x 17 conditions). The experiment was divided into 14 sessions of 10 minutes each during which all the conditions were randomly intermingled with an inter-trial-interval of 0.9s. Visual-only, vestibular-only and bimodal stimuli were therefore alternated on a trial-by-trial basis and predicting the nature of the stimulus was impossible. The subjects took regular breaks between sessions and completed the entire experiment in approximately 4 hours. The direction of rotation (left or right) was randomly chosen on each trial.

To test for mandatory fusion, subjects had to pick out the odd stimulus among three successive rotations in a three-alternative forced choice task (Fig 2c). The same standard rotation size and the same trial timings were used as in the first experiment. The test rotation sizes were seven equally spaced angles in the interval $9°$-$21°$. The sizes were again chosen based on preliminary tests conducted on two subjects and such that they would include one point on each end of the fitting curve where size discrimination could be achieved with nearly 100% certainty. Two of the rotations were the test and one the standard (or vice versa) and were presented in
random order. For bimodal stimuli the test rotation included conflicting visual and
vestibular cue pairs \((S_{vi} = S + \Delta_{vi}, S_{ve} = S + \Delta_{ve})\). We tested eight conditions of
bimodal pairings, each corresponding to a different \(\Delta_{vi}/\Delta_{ve}\) ratio (1, 0.5, 2, 0, \(\infty\), -1, -
0.5, and -2) (see Results for details). The different ratios yield different amounts of
conflict between \(S_{vi}\) and \(S_{ve}\), with no conflict in the 1 condition and maximum conflict
in the -1 condition. For the 1 ratio, \(S_{vi}\) and \(S_{ve}\) were always equal and therefore
corresponds to the easiest condition for picking out the odd stimulus (tested \(\Delta_{vi}\) and
\(\Delta_{ve}\) values were 0, ±1.71, ±3.43 and ±5.14). For the -1 ratio, the average of \(S_{vi}\) and
\(S_{ve}\) was always equal to the standard rotation size \(S\) and, in the case of equal cue
reliabilities, makes the identification of the odd stimulus theoretically impossible if
single cues remain inaccessible. For the \(\infty\) ratio \(\Delta_{ve} = 0\), meaning that \(S_{ve}\) was always
equal to the standard rotation size while \(S_{vi}\) took on the seven equally spaced angles
in the interval 9°-21°. The opposite was true for the 0 condition. For 2 and -2 ratios,
tested \(\Delta_{vi}\) values were 0, ±1.71, ±3.43 and ±5.14 and the \(\Delta_{ve}\) values were calculated
from the \(\Delta_{vi}/\Delta_{ve}\) ratio. The opposite was true for the -0.5 and 0.5 ratios. The subject
had to answer whether the first, second or third rotation was different than the other
two on any basis. Each comparison was repeated 10 times across the 10 conditions
(2 unimodal and 8 bimodal) producing a total of 700 trials per subject, randomly
intermixed in 10 sessions of 10 minutes each. In both experiments, subjects initially
underwent multiple short training sessions to familiarize themselves with the task and
to ensure better than chance performance. Data collected on these sessions were
not used for analysis.

Data analysis
All data analyses were performed offline with custom programs compiled in MATLAB (Mathworks). For each test angle, individual answers were pooled across all subjects to obtain a probabilistic measure of the response and yield a sufficient sample set for the statistical comparisons. It consisted of calculating the proportion of “bigger” (Fig. 3) or “incorrect” (Fig. 6) responses based on 80 answers (10 from each subject) for every test angle. Using the method of least squares, the proportion of “bigger” responses in the first experiment were fit with a cumulative Gaussian function and the proportion of incorrect responses in the second experiment were fit with a Gaussian. Measures of the mean, variance and discrimination threshold were then extracted from the obtained fits in each condition. A bootstrap analysis provided standard errors for each measure and allowed a statistical comparison between the experimentally measured values and model predictions. It consisted of repeating the data fit for each condition 9999 times on a different subset of responses each time. The different subsets were formed by taking at random, with replacement, N trials from the total set of N for each test angle (for N=80, $10^{23}$ such combinations are possible). The standard deviation of 9999 repeated measures is then the standard error of the measure obtained using the original data set. Statistical tests were made by assessing the amount of overlap between the bootstrap iterations of two measures. If the measure of interest is $\sigma$ and, $\sigma_{ex}^j$ and $\sigma_{pr}^j$ are its experimental and predicted estimates obtained from the $j^{th}$ bootstrap sample, then the one-tailed bootstrap probability of $(\sigma_{ex} > \sigma_{pr})$ is

$$p = \frac{1}{B} \sum_{j=1}^{B} I(\sigma_{ex}^j - \sigma_{pr}^j > 0)$$
where B=9999 and I() is the indicator function, which is equal to 1 when its argument is true and 0 otherwise. The inequality would be reversed for the probability of $(\sigma_{ex}^j < \sigma_{pr}^j)$. The one-tailed bootstrap p value is therefore simply the proportion of $(\sigma_{ex}^j - \sigma_{pr}^j)$ values that are more extreme than 0. We prefer this approach to parametric testing because it provides a direct computation of the cumulative distribution of a test statistic instead of having to use an asymptotic approximation.

Results

Optimal integration of visual and vestibular cues

In order to experimentally obtain the visual and vestibular likelihoods and the posterior resulting from their integration when perceiving own-body rotations, subjects made relative judgments about the size of two consecutive rotations; a standard rotation of fixed angular displacement and a test rotation of variable size (Fig. 2c and Materials and Methods). The probabilities of perceiving the test rotation as bigger than the standard were fit by a cumulative Gaussian function for the tested range of angles and the Gaussian likelihoods and posterior can be derived for each condition by taking the mathematical derivative of the fits.

Variance measures extracted from the experimentally obtained sensory likelihoods were used to predict the variance of the posteriors according to MLE and Equation 2 (see Materials and Methods). Because these were evaluated by comparing the size of two successive rotations, they actually correspond to the true likelihood variances scaled by a factor of 2. The variances measured experimentally using the bimodal stimuli (mean±sem: $\sigma_p=2.7±0.2, 3.1±0.24, 3.3±0.26, 3.9±0.35$ for
the four coherence levels) were reduced relative to using either single cue alone ($\sigma_{ve}=5.1\pm0.54; \sigma_{vi}=3.5\pm0.28, 3.7\pm0.3, 5.0\pm0.53, 7.3\pm1.1$) and closely matched the predictions for all four visual coherence levels (Fig. 3a). The result was readily reproducible across individual participants. The standard deviations of the psychometric fits to the bimodal data of individual subjects across coherence levels were not statistically different between the measured and MLE predicted values (paired t-test, $p=0.56$) and are shown in Figure 4a. The same bimodal thresholds were however significantly reduced relative to the smallest single cue thresholds (paired t-test, $p<0.001$) confirming that integrating information from visual and vestibular senses reduces perceptual uncertainty in a statistically optimal manner when estimating rotatory self-motion. As the critical test for demonstrating cue integration we furthermore provide a comparison between the bimodal and single-cue thresholds for each subject in the case where the visual and vestibular weights were closest to being equal (Fig. 4b).

To test for the predictions of Equation 3, size comparisons were made against a standard stimulus in which the chair and the visual scene rotated by different amounts, so that $\hat{S}_{vi} \neq \hat{S}_{ve}$. When $\hat{S}_{ve}$ signaled a bigger rotation than $\hat{S}_{vi}$, the subjects’ most likely posterior estimate $\hat{S}_p$ increased with decreasing visual coherence level (black trace in Figure 3b). The opposite was observed when $\hat{S}_{vi} > \hat{S}_{ve}$ (blue trace in Figure 3b). This is what is predicted if the subjects dynamically attribute more weight to the vestibular cue than the visual according to Equations 3 and 4, as the reliability of the latter is reduced (dotted lines in Figure 3b). Pooling the data from the two conflict conditions and expressing it in terms of visual and vestibular weights (Fig. 3c), shows that cue reweighing occurs, follows the MLE predicted trend, but deviates
from optimality because subjects tended to significantly overweigh the visual cue. The extent of the visual bias was however variable across individuals (Fig. 4c).

*Mandatory fusion*

We next addressed the issue of mandatory cue fusion by adopting a paradigm (Hillis et al. 2002) measuring an observer’s ability to perceive two rotations as being different. Subjects were asked to detect the odd stimulus among three successive rotations, two of which were the standard and one the test (or vice versa) (Fig. 2c and Materials and Methods). This oddity task is advantageous to outright asking the subjects whether two rotations are different, because it forces them to adopt the same decision criterion and thus eliminates response bias from the observer’s actual discriminability (Swets 1961) (i.e. the subjects stay unaware of the task’s real purpose; perceptual responses are not contaminated by higher level cognitive strategies).

In the unimodal cases, comparisons were made between a standard rotation of size $S$ and a test of size $S + \Delta$. The proportions of incorrect responses for each $\Delta$ were fit by a Gaussian function (Fig. 5a). The fits symbolize the distribution over trials of visual and vestibular estimators $\hat{P}_{vi}$ and $\hat{P}_{ve}$, the probabilistic descriptions of the observer’s discrimination ability using each sense alone, and the associated single cue discrimination thresholds $T_{vi}$ and $T_{ve}$ that we define here as equal to one standard deviation of the Gaussian fits. The coherence level used for the visual stimulus was the one that yielded $\sigma_{vi} \approx \sigma_{ve}$ for the average subject when optimal integration was tested in the first experiment (see Fig. 3a). In the bimodal case, discriminability between the pairs $(S_{vi} = S, S_{ve} = S)$ and $(S_{vi} = S + \Delta_{vi}, S_{ve} = S + \Delta_{ve})$ was assessed. We tested eight conditions of bimodal pairings, each corresponding to
a different $\Delta_{vi}/\Delta_{ve}$ ratio. A ratio of 1 generates visual and vestibular stimuli that always rotate by the same amount and the bi-sensory estimator $\hat{p}_{bi}$ should provide better discrimination ability than either single cue estimator in that case (Fig. 1c). A ratio of -1 produces conflicting visual and vestibular cues but identical bi-sensory estimates between the standard and the test according to Equation 3 when $w_{vi} \approx w_{ve}$. Theoretically, the ability to discriminate the two metameric stimuli will be compromised if the observer only uses $\hat{p}_{bi}$ and does not retain the unimodal estimators $\hat{p}_{vi}$ and $\hat{p}_{ve}$ (Fig. 1d). The remaining six conditions corresponded to $\Delta_{vi}/\Delta_{ve}$ values (0.5, 2, 0, $\infty$, -0.5, -2) that produce bimodal stimuli with varying amounts of cue conflict that lie between the latter two extremes.

If all three estimators are accessible, in the cue space depicted in Figure 5b, the predicted discrimination thresholds (blue dots) would lie on the contour defined by the red (Equations 5 and 6) and green (Equation 7) lines. That is if the subjects do not make an additive use of all three estimators. Probability summation is actually unrealistic because it assumes that the three estimators can be used independently. Given the fact that $\hat{p}_{bi}$ is a weighted average of $\hat{p}_{vi}$ and $\hat{p}_{ve}$ the independence assumption is clearly invalid. The probability summation model will thus not be considered further and in any case makes a less conservative prediction than the model in Figure 5b. If mandatory fusion was to occur, perceptual access to $\hat{p}_{vi}$ and $\hat{p}_{ve}$ would be lost and discrimination performance would uniquely be based on the bi-sensory estimator. $(S, S)$ could be perceived as different from $(S + \Delta_{vi}, S + \Delta_{ve})$ only if Equation 7 is satisfied, giving rise to the theoretical prediction shown in Figure 5c.

The prediction depicted in Figure 5b was compared against the mandatory fusion prediction (Fig. 5c) to see which best accounts for the discrimination
thresholds obtained experimentally in the eight tested conditions. The predicted contours were constructed using the measured $T_{vi}$ and $T_{ve}$ (mean±sem: 4.54±0.42 and 4.74±0.45, respectively), and the estimated $T_{bi}$. Our results show that the subjects’ discrimination thresholds (again defined as equal to one standard deviation of the obtained Gaussian fits) for perceiving $(S, S)$ as different from $(S + \Delta_{vi}, S + \Delta_{ve})$ were in many cases larger than those predicted from an independent use of all three estimators (Fig. 6a). Individual subjects indeed showed consistent losses in discrimination performance when compared to the prediction of using single cue estimators, even without adding probabilistic responses (Fig. 6b). Threshold predictions were derived for each subject individually from the corresponding measured single cue thresholds. For $\Delta_{vi}/\Delta_{ve}$ ratios (0.5, 1 and 2) where using $\hat{P}_{bi}$ leads to a gain in performance over using $\hat{P}_{vi}$ and $\hat{P}_{ve}$, single subject thresholds (Fig. 6b) were not statistically different ($p>0.05$, t-test) from the identical predictions of Figures 5b and 5c. On the other hand, in the remaining conditions where mandatory fusion leads to a worse performance, subjects showed losses in discrimination ability (Fig. 6b) relative to using the single cue estimators ($p<0.05$, for $\Delta_{vi}/\Delta_{ve} = -2$, -1, -0.5 and 0, average power=0.81, and $p>0.3$ for $\infty$, one-tailed t-test). The results are therefore best accounted for by assuming that only the bi-sensory estimator stemming from sensory fusion is available for making perceptual judgments.

Conditions where very large, even infinite, discrimination thresholds are predicted ($\Delta_{vi}/\Delta_{ve} = -0.5, -1, -2$) are those that involve stimuli with the largest conflict angles between the visual and vestibular cues. Because the subjects were free to detect the odd stimulus on any basis, it is reasonable to assume that detection was based on sometimes perceiving the conflict in trials corresponding to the biggest $\Delta$ values. The absence of optimal cue integration in those cases explains the non-infinite or lower
than theoretically expected thresholds. Also, a perceptual aspect other than rotation size or conflict detection might have transpired in cases of large conflict and been used by the subjects to do the task. We finally note that visuo-vestibular integration resulted in overweighing of the visual cue (Fig. 3c): including a visual bias in the model would increase $w_{vi}$ and decrease $w_{vc}$ in Equation 7, translate into the green lines being rotated clockwise in Figure 6b and therefore yield an even better correspondence between experimental results and the mandatory fusion prediction.

**Discussion**

When perceiving properties of the world observers combine redundant information from different sensory modalities. Perceptual estimates derived from each of these sensory cues will naturally exhibit variability across repeated observations of the same stimulus. It has been repeatedly shown that the variability is reduced in a statistically optimal manner if cues are combined to produce the final percept. The neural mechanism underlying multisensory integration is therefore likely to be a process of probabilistic inference seeking to reduce perceptual uncertainty (Knill and Pouget 2004). Our results demonstrate, and extend what has recently been reported for the perception of heading (Butler et al. 2010; Fetsch et al. 2009), that the central nervous system integrates multisensory idiothetic information according to the same laws of probability. Indeed, the same optimal reduction of variance and cue reweighting in proportion to relative reliability occurs as for non-idiothetic cues (Fig. 3). When perceiving whole body rotations, subjects however tended to overweigh the visual cue (Fig. 3c and Fig. 4c), which is in contrast with heading perception where an excessive reliance on the vestibular cue is observed (Butler et al. 2010; Fetsch et al. 2009). A selective over-weighing of otholith signals
and an under-weighing of semi-circular canal signals might thus be occurring when integrated with vision. Similarly suboptimal cue weights have been observed for other sensory modalities as well (Battaglia et al. 2003; Knill and Saunders 2003), and why such deviations from optimality occur is not fully understood. One possibility is that the biases might be reflective of a recalibration mechanism, where one estimate is remapped and realigned with the other in an attempt to internally correct the inconsistencies in multisensory input. Such a process is indeed known to occur when the observer is constrained to integrate conflicting information (Adams et al. 2001; Block and Bastian 2011), even after only a very short exposure (in the milliseconds range) to the conflict (Wozny and Shams 2011).

When sensory cues are combined, perception is governed by the posterior; the product of their integration. As they are individually less reliable, neural signals underlying the two isolated sensory likelihoods are therefore not used at the perceptual level. It has been demonstrated that when the two cues originate from different sensory modalities (vision and touch for example), observers still have access to this unused unimodal information (Hillis et al. 2002). Why might the ability to access these “useless” signals be preserved? In laboratory settings, experimental conditions can be constructed where using individual likelihoods, if available, is advantageous in terms of better task performance, even in the presence of a more reliable estimate (Fig. 1d). Such conditions, however, have little ecological validity. They can only serve as a useful experimental tool for testing whether likelihood signals are accessible but do not provide a valid explanation for their accessibility. The reasons might instead be rooted in the causal inference process (Koerding et al. 2007; Parise et al. 2012; Shams and Beierholm 2010) as observers often effortlessly attribute separate causes to simultaneously received cross-modal cues. The
dissociated cues are not integrated and each gives rise to an independent percept. Preserving unused information when cues are integrated might thus be reflective of a neural organization sculptured by the experience that most cues simply do not have the same cause in everyday life. Actually, the MLE model of Equation 1 is a simplification, under the common cause assumption, of a more generalized hierarchical causal inference (HCI) model (Shams and Beierholm 2010). In the full HCI model, the a posteriori estimate is a weighted average of the MLE estimate of Equation 3 and one of the single cue estimates, the one that dominates when separate causes are inferred. The weighting is determined by the probability of the common cause scenario. Therefore, whenever the latter probability is not equal to one, the individual likelihoods must remain accessible for one of them to be combined with the product of their initial integration.

Our results show that this generalization does not apply to cross-modal integration when the object of perception is the perceivers own body. Visual and vestibular idiothetic cues are individually discarded after fusing into a single percept (Fig. 6) similarly to non-idiothetic cues within the same sense. Indeed, mandatory fusion has so far only been observed for the integration of unimodal cues; binocular disparity and texture gradients when perceiving the slant of a surface (Hillis et al. 2002; Nardini et al. 2010). These two cases share the characteristic of cues not being dissociable in natural conditions; they are necessarily redundant and always integrated. The probability of the common cause scenario is therefore never different from unity. Even when observers are instructed to actively attend either the visual or vestibular stimulus and ignore the other, providing the cue-conflicts go unnoticed, they seem incapable of weighing the two cues independently (Berger and Bueltthoff 2009). Because there is a cost, in terms of brain resources, associated with carrying
two neural representations of the same rotation stimulus, mandatory fusion likely
results from the absence of an evolutionary pressure to preserve the individual
channels for stimuli including both information sources. Sensory signals for
perceiving self-motion thus seem to be processed more like unimodal cues rather
than as originating from separate senses. Additionally, the particular rotations we
used (Fig. 2b) are faithfully encoded by both visual and vestibular systems. For a
broader range of stimuli however, both systems are individually deficient. Low-
frequency, low-acceleration vestibular stimuli and high-frequency, high-acceleration
visual stimuli are inaccurately encoded at early stages of neural processing (Waespe
and Henn 1977; 1979). Only neural signals generated by rotations including both
sensory inputs give a truthful account of the actual motion of the body over the entire
operating range (Dichgans et al. 1973; Waespe and Henn 1977; 1979). A loss of
inaccurate uni-sensory information might be incurred for the benefit of a more
accurate bi-sensory signal, thus providing a potential neurophysiological basis for the
mandatory fusion that we observed. We postulate that the same is also likely to be
ture for heading perception (Butler et al. 2010; Fetsch et al. 2009) but not in other
stances where optimal cross-modal cue integration has been previously
demonstrated (Alais and Burr 2004; Ernst and Banks 2002; Mendonca et al. 2011).
This also applies to own hand position perception using visual and proprioceptive
cues (van Beers et al. 1999; vanBeers et al. 1996), despite the idiothetic nature of
the latter, because it is identical, in all things that matter, to estimating the position of
an external object held in that hand.

It follows that passive whole body rotations can be simulated by combinations
of visuo-vestibular stimuli that, like color metamers, differ physically but "look" the
same. Our result is also akin to examples of visual metamers, as for instance two
verniers with opposite offsets flashed in quick succession that are fused and become indistinguishable from a single almost aligned vernier (Scharnowski et al. 2009). Color metamers occur because spectral information is lost by cone photoreceptors in the retina (Wandell 1995) and therefore cannot be recovered in the brain. In the case of visual and visuo-vestibular metamers, information loss necessarily occurs at a neural level upstream of where the metameric stimuli would yield identical neural responses. We envision that our findings might guide future electrophysiology and modeling approaches to elucidate where and how visuo-vestibular metamers are formed, similarly to recent attempts at explaining visual metamerism (Freeman and Simoncelli 2011) and at examining the dynamics of the fusion process (Scharnowski et al. 2009). Finally, mandatory fusion of visual and vestibular signals can explain why neurons in cortical and subcortical vestibular centers are highly multi-modal (Akbarian et al. 1988; Bremmer et al. 2002; Buttner and Buettner 1978; Buttner and Henn 1976; Dichgans et al. 1973; Duffy 1998; Grusser et al. 1990; Henn et al. 1974; Meng and Angelaki 2010; Page and Duffy 2003; Schlack et al. 2002; Takahashi et al. 2007; Waespe et al. 1981; Waespe and Henn 1977; 1979), most often visually-responsive, and hints at developmental reasons for the absence of a centralized vestibular cortex (Fukushima 1997; Guldin and Grusser 1998).
Acknowledgments

We thank Bruno Herbelin for technical assistance and Danilo Jimenez Rezende for advice on data analysis, as well as Michael Herzog and Wulfram Gerstner for comments on an earlier version of the manuscript. This study was supported by the Swiss National Science Foundation (SINERGIA CRSII1-125135/1).

References


Henn V, Young LR, and Finley C. Vestibular nucleus units in alert monkeys are also influenced by moving visual-fields. *Brain Res* 71: 144-149, 1974.


van Beers RJ, Sittig AC, and van der Gon JJD. Localization of a seen finger is based exclusively on proprioception and on vision of the finger. Experimental Brain Research 125: 43-49, 1999.


Figure 1. Predictions of the optimal observer framework. \textit{a, b}, A statistically optimal integration of sensory likelihoods results in a reduction of perceptual uncertainty according to Equation 2 (\textit{a}) and a reliability based cue reweighting according to Equation 3 (\textit{b}). \textit{c, d}, For rotations comprised of both stimuli, optimal integration of the visual and vestibular likelihoods results in the posterior estimate according to Eq. 1. Discrimination between a standard and a test rotation can be based on either the single cue estimators $\hat{p}_{vi}$ and $\hat{p}_{ve}$ (light and dark grey) or the bi-sensory estimator $\hat{p}_{bi}$ (black trace). Traces symbolize here the distribution of the estimates over trials. Successful discrimination is achieved when the physical difference between the estimates of the two rotations ($\Delta_{vi}$, $\Delta_{ve}$ and $\Delta_{bi}$) is greater than the respective discrimination thresholds (defined here as $\sigma_{vi}$, $\sigma_{ve}$ and $\sigma_{bi}$). Using the bi-sensory estimator can either improve discriminability as in \textit{c} or not, as in \textit{d}. Empirical discrimination thresholds can therefore reveal which estimators the subjects have access to for making perceptual judgments. See Materials and Methods for definition of all symbols and terms.

Figure 2. Experimental setup and paradigms. \textit{a}, Schematic of the vestibular and visual stimuli delivered to the subjects for perceiving whole body yaw rotations. \textit{b}, Position and velocity profiles of the rotation stimulus. \textit{c}, Trial timeline of the optimal integration paradigm used to determine the sensory likelihoods and posterior (top). Subjects made size comparisons between a standard and a test stimulus. Visual and vestibular only conditions involved stimuli within a single modality. In the bimodal condition, the conflict angle $\Delta$ between the visual and vestibular rotations was 0°, 4° or -4°. Trial timeline in the oddity detection task used to test for mandatory fusion.
Subjects had to detect the odd stimulus among three successive rotations. The odd stimulus could either be the standard or the test, and come either first, second or third.

Figure 3. Optimal integration of visual and vestibular cues. a, Standard deviations obtained from the measured visual, vestibular and bi-sensory likelihoods for four different levels of visual coherence. The empirical bi-sensory values were not significantly different (n.s.: p>0.2, one-tailed bootstrap test) from the MLE predictions. b, Maximum likelihood estimates extracted from the measured posterior distributions and compared to the MLE predictions for the two conflicting visual and vestibular stimuli (S_v=13°, S_vs=17°) and (S_v=17°, S_vs=13°). c, Averaged data from the two conditions in b and expressed in terms of weights show a suboptimal cue reweighting when compared to model predictions (*: p<0.05, **: p<0.01, one-tailed bootstrap test). All delimiters correspond to bootstrap standard error.

Figure 4. Individual subject data. a, Similarity between the predicted and measured standard deviations of the posterior across all subjects and for all four coherence levels. Colors denote different subjects and symbols correspond to the different coherence levels. b, Difference between the lowest single-cue standard deviation and the measured (blue bars) and predicted (red lines) bimodal standard deviations for the coherence level yielding best matched visual and vestibular weights for each subject. c, Difference between predicted and empirical visual weights averaged across the four coherence levels and sorted in descending order for the eight subjects.
**Figure 5.** Theoretical predictions for rotation size discrimination. **a,** Mean responses for choosing the odd visual or vestibular stimulus (i.e. perceiving $S + \Delta$ as different than $S$) obtained by pooling raw choices from all subjects. The Gaussian fits symbolize the visual and vestibular estimators and provide the unimodal discrimination thresholds $T_{vi}$ and $T_{ve}$. **b,** Cue space where each point corresponds to a bimodal pairing (abscissas and ordinates indicate rotation sizes signaled by the visual and vestibular stimuli, respectively). When perceiving $(S + \Delta_{vi}, S + \Delta_{ve})$ as different than $(S, S)$, if the cues are integrated without mandatory fusion, discrimination thresholds (blue dots) are predicted to lie on the contour defined by the red (Equations 5 and 6) and green (Equation 7) lines. Gray lines and main axes each correspond to a $\Delta_{vi}/\Delta_{ve}$ value representing one of the tested conditions. The yellow area comprises data points theoretically indistinguishable from $(S, S)$.

**Figure 6.** Comparison between predicted and measured discrimination thresholds. **a,** Compared to single cue thresholds (red), pooled data from all subjects showed improved discrimination ability (“Gain”) when using the bi-sensory estimator (green) is advantageous (0.5, 1 and 2 conditions) and worse discrimination (“Loss”) when it is not ($\infty$, -2, -1, -0.5 and 0 conditions). **b,** Loss or gain in discrimination ability compared to the theoretical prediction of Figure 5b for individual subject data as measured by the $d_{ex}/d_{th}$ ratio. $d_{ex}$ is the distance between the cue space origin and experimental data (blue), and $d_{th}$ between the origin and the theoretical limits (red or green), as depicted in **a.** All shadings and delimiters correspond to bootstrap standard error.
Optimal integration paradigm

<table>
<thead>
<tr>
<th>Standard rotation</th>
<th>Test rotation</th>
<th>Bigger?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{vi}=15^\circ+\Delta_{vi}$</td>
<td>$10^\circ&lt;S_{vi}&lt;20^\circ$</td>
<td>(Yes/No)</td>
</tr>
<tr>
<td>$S_{ve}=15^\circ-\Delta_{ve}$</td>
<td>$10^\circ&lt;S_{ve}&lt;20^\circ$</td>
<td></td>
</tr>
</tbody>
</table>

Mandatory fusion paradigm

<table>
<thead>
<tr>
<th>Standard rotation</th>
<th>Test rotation</th>
<th>Standard rotation</th>
<th>Odd rotation? (1st/2nd/3rd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{vi}=15^\circ$</td>
<td>$S_{vi}=15^\circ+\Delta_{vi}$</td>
<td>$S_{vi}=15^\circ$</td>
<td></td>
</tr>
<tr>
<td>$S_{ve}=15^\circ$</td>
<td>$S_{ve}=15^\circ+\Delta_{ve}$</td>
<td>$S_{ve}=15^\circ$</td>
<td></td>
</tr>
<tr>
<td>0 0.5 1.8 3.5 4 6</td>
<td>0 0.5 1.8 3.5 4 5.3 5.8 7.8</td>
<td>0 0.5 1.8 3.5 4 5.3 5.8 7.8</td>
<td></td>
</tr>
</tbody>
</table>