Learning to Combine Arbitrary Signals from Vision and Touch

Frank Jäkel[1] and Marc O. Ernst[2]

[1]Graduate School for Neural and Behavioural Sciences, Tübingen, Germany.
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Abstract. When different perceptual signals of the same physical property are integrated–e.g., the size of an object, which can be seen and felt–they form a more reliable sensory estimate [3]. This however implies that the sensory system already knows which signals belong together and how they are related. In a Bayesian model of cue integration this prior knowledge can be made explicit. Here, we examine whether such a relationship between two arbitrary sensory signals from vision and touch can be learned from their statistical co-occurrence such that they become integrated. In the Bayesian model this means changing the prior distribution over the stimuli. To this end, we trained subjects with stimuli that are usually uncorrelated in the world–the luminance of an object (visual signal) and its stiffness (haptic signal). In the training phase we presented only combinations of these signals which were highly correlated. Before and after training we measured discrimination performance with distributions of stimuli which were either congruent with the correlation during training or incongruent. The incongruent stimuli came from an anti-correlated distribution compared to the stimuli during training. If subjects were sensitive to the correlation between the signals then we would expect to see a change in their prior knowledge about what combinations of stimuli are usually encountered. Accordingly, this should change their discrimination performance between pre- and post-test. We found a significant interaction between the two factors pre/post-test and congruent/incongruent. After training, discrimination thresholds for the incongruent stimuli are increased relative to the thresholds for congruent stimuli, suggesting that subjects learned to combine the two signals effectively.

1 Introduction

Our brain constantly receives sensory information from many different sources and modalities. Some of these sensory signals have to be combined in order to form a coherent percept of the world–others have to be kept separate. For example, when moving your head visual, vestibular and proprioceptive signals all give rise to an estimate of the position and orientation of your head. Hence, it would make sense for the sensory system to integrate these signals into one representation of head position and orientation. Another prominent example is depth perception. The distance to an object can be estimated from the disparity signal between the two eyes' images, relative size, perspective, motion parallax, and other cues [1, 6]. When touching an object there are also several signals that can be used to judge its size when it is simultaneously explored by vision and touch [3]. All these sensory signals, which are at least partially redundant when derived from the same object property or event, should be combined by our nervous system. Other signals, which are not elicited by the same object property or event, should not be combined. It is the job of our brain to decide which cues to combine, and which not. The question we address here is whether we can learn to combine different signals which are usually not combined. Phrased differently, we ask whether the combination of signals is pre-determined and hard wired in the nervous system or whether it is adaptive.

Each sensory signal on its own is inherently noisy. The advantage of combining sensory information from different sources is that the noise in the combined signal is reduced relative to each signal alone [2, 3, 7, 10]. An optimal model of cue combination describing this benefit can easily be derived under the assumption that the noises are Gaussian and independent. Let's call the stimulus a subject is presented with S . The estimator with the lowest possible variance σ^2 can be derived from the maximum likelihood principle. This combined estimate $\hat{s}$ is given by the weighted sum of the individual estimates $\hat{s}_i$ based on the individual signals, with weights w_i proportional to their reciprocal variances σ_i^2 :

$$\hat{s} = \sum_i w_i \hat{s}_i \qquad \text{with} \qquad w_i = \frac{\sigma_i^{-2}}{\sum_j \sigma_j^{-2}} \ . \tag{1}$$

With this particular choice of weights the variance of the combined estimate σ^2 becomes minimal and is given by

$$\sigma^2 = (\textstyle\sum_i \sigma_i^{-2})^{-1} \ . \tag{2}$$

It was recently shown by several studies that the human brain integrates sensory information in such an optimal way [3, 5, 9]

How does the brain know which signals can be combined? When seeing and feeling small objects they usually look small but also feel small whereas big objects look and feel big. This trivial statement demonstrates that there is a natural statistical relationship between the felt and seen size of an object. This statistical relationship could have been exploited by the developing brain in order to form its own integrated concept of the objects' property now called size. Here we want to test this idea and see whether our brain can adapt to integrate two signals, which are usually not combined because they are signaling two different properties of an object. We decided to test this in a framework of visual-haptic integration. We chose the luminance of an object as the visual dimension and its stiffness as the haptic dimension. This choice was made because we believe there should be no statistical relationship between these two properties in a "natural" environment. By correlating these properties (and hence the visual and haptic signals derived from them) in an artificial environment we can introduce a new statistical relationship between them. A comparison of discrimination performance before and after extensive training with such correlated stimuli will reveal whether cue combination can be learned from the statistical co-occurrence of the stimuli in the environment.

Bayesian estimation theory provides a principled approach to handle such questions. In the following we will introduce such a model based on maximum a posteriori (MAP) estimation. As opposed to the maximum likelihood (ML) approach outlined above, the maximum a posteriori formulation can use beliefs about the stimulus statistics (top-down information) in the form of priors. Learning in this view is reflected in a change of the prior beliefs about the distribution of the stimuli.

1.1 Combination of Signals as MAP estimation

From an "ideal observer" perspective the problem of combining different sources of information is conveniently addressed using probability theory [2, 8, 11, 12, 15]. The physical stimulus we used has a visual (luminance) and haptic (stiffness) property–s_V and s_H –that results in a visual and haptic sensory signal (l_V and l_H). Faced with the sensory signals how does the brain infer the underlying physical stimuli? The Bayesian approach explicitly uses prior beliefs about the distribution of the stimuli. Given that the sensory noise in the visual and haptic channel is distributed normally the generation of the sensory signals is modeled with a 2D Gaussian about the physical stimulus:

$$p(l \mid s, L) = N(l; s, L) \tag{3}$$

with a physical visual/haptic stimulus $s = (s_V, s_H)$, a sensory signal $l = (l_V, l_H)$ and the covariance matrix L which is positive definite and symmetric. Interpreted as a function of s this is called the likelihood of s given the sensory signal l. A graphical illustration of such a likelihood function is depicted in Fig. 1 (top row).

In the next step we model the prior distribution which describes what stimuli are expected to be seen by the subject. It is noteworthy to point out that this does not necessarily mean that the stimuli are really drawn from this distribution. The prior is supposed to model what the subject expects to see, irrespective of what the real distribution is. Of course, the prior information can be misleading if the stimuli presented don't come from the distribution assumed by the subject. Recently, this fact has been taken to explain several sensory illusions [12]. Here, for simplicity and illustration we assume the prior to be also Gaussian:

$$p(s \mid p, P) = N(s; p, P) \tag{4}$$

with a mean p and a covariance matrix P. Some schematics for prior distributions are given in the middle row of Fig. 1.

According to Bayes' rule we can calculate the posterior distribution over the stimuli s given the subject's observations l and the subject's beliefs present in the prior

$$p(s \mid l, L, p, P) = \frac{p(l \mid s, L) \cdot p(s \mid p, P)}{p(l \mid L, p, P)} = N(s; \hat{s}, S) \ . \tag{5}$$

In the Bayesian view this should now represent the subject's belief about what stimulus he has encountered. Fortunately this posterior is again a Gaussian with a covariance matrix S

$$S = (L^{-1} + P^{-1})^{-1} \tag{6}$$

and a mean $\hat{s}$ which corresponds to the weighted average of likelihood and prior (see lower row of Fig. 1 for an illustration of the posterior):

$$\hat{s} = W_L l + W_P p = S(L^{-1} l + P^{-1} p) \; . \tag{7}$$

The maximum of this distribution, which is equal to the mean $\hat{s}$, is taken to be the estimator for the presented stimulus *s*. This estimate is called the maximum a posteriori (MAP) estimate.

The MAP estimator can be thought of as the optimal way to incorporate extra-sensory knowledge about the stimulus distribution. Where the likelihood function (given by *l* and *L*) describes the noisy sensory signals (bottom-up if you want), the prior distribution (given by *p* and *P*) describes the extra-sensory (or top-down) prior beliefs that the subject has, such as knowledge about the correlation between the physical stimuli.

If the subject knows nothing about the distribution of the stimuli, i.e. there is no prior belief (informally, $P \rightarrow \infty$) about the distribution of the stimuli, the terms for the prior disappear and only the likelihood function (i.e., the sensory signal) is left, so that the MAP estimator will simply be the sensory signal *l* (flat prior: left column in Fig. 1). If the variances are finite then they will pull the estimator towards the mean of the prior *p* (middle column in Fig. 1). The smaller the variance of the prior in a given direction the more the estimate is drawn to the center of the prior. An extreme case is where the variance in one direction becomes zero (right column of Fig. 1). This means it is certain that the stimulus does not change along this direction and the estimator is equal to the center of the prior along this direction. This corresponds to a situation of complete fusion of the two signals. Another extreme case is the situation when the noise in the sensory signals goes to infinity (Eigenvalues of *L* go to infinity) so there is basically no useful sensory information. Decisions then have to be based only on prior knowledge (this scenario is not depicted in Fig. 1).

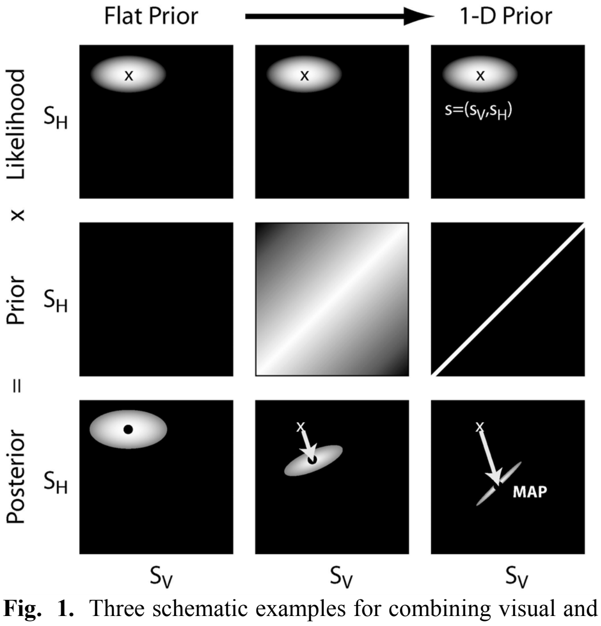

Fig. 1. Three schematic examples for combining visual and haptic signals with different priors (columns). *Top row:* Likelihood distributions; (x) physical stimulus. *Middle Row:* Prior distributions; left: flat prior, middle: infinite variance in one direction and a variance of one perpendicular to that, right: again infinite variance in one direction but this time zero variance perpendicular to that. *Lower Row:* Posterior distributions which are the product of the likelihood and prior distributions. The MAP estimate is indicated by the (•).

1.2 Learning to Combine Signals

How is learning to combine signals represented in the MAP model? For the experiment we suggest it is natural to think about learning as a change of the subject's belief about the distribution of stimuli, which is reflected in a modification of the priors. It is natural to think about it as a change of priors because it is the joint distribution of the stimuli we are manipulating in the experiment: before learning visual and haptic properties were uncorrelated and during learning they are being correlated. Most other adaptation paradigms treat learning as a modification somewhere in the nervous system, i.e. the map from the physical stimulus to the sensory signal is altered. This map from the physical scene to the sensory signal is given by the likelihood function in our framework. For us the likelihood simply describes the inherent physical noise in the sensory system of an agent. Although we cannot exclude that the noise, and hence the likelihood function, changes over the course of the experiment due to perceptual learning, for the time being we will assume that it does not and focus on the change of the beliefs about the manipulated distributions. By changing the stimulus distribution we effectively try to manipulate subjects' assumptions about the environment. By using Bayes' rule a subject could infer its state of belief about the world taking into account the sensory data, the internal noise and its prior beliefs about the environment. Bayes' rule is, so to say, the ecological version of Helmholtz's "unconscious inference".

This can be illustrated with two stimuli, which we assume to be independent, for example the luminance of an object and its stiffness. Say, whether an object reflects much light or not does not tell anything about how hard or soft it feels. Therefore, in a "natural" environment it does not make sense to combine the sensory signals elicited by these two stimuli and a subject should belief that they are independent of each other. However, if we lived in a world where bright objects always feel hard and dark objects soft, it would suddenly make sense for our sensory system to combine the visual and the haptic signal. In other words, if the value of one variable is informative about the value of the other and the system knows their joint distribution then it would be useful to combine these signals.

Coming from a world where luminance and stiffness are independent and then put into a situation where they are highly correlated (i.e., not independent), what is changing? Learning to combine these two signals means a change in belief about their joint distribution, which in the Bayesian framework is equal to a change in the prior distribution.

To test this hypothesis we measured subjects' discrimination performance for objects that can vary in luminance and stiffness before and after extensive training with stimuli for which these two properties were highly correlated–we even set the correlation to one. Given the MAP model we can make some qualitative predictions of how the discrimination performance should change ideally. The predictions are only qualitative because it is very hard to exactly match all the assumptions present in the model in an experiment. However, this MAP analysis we presented above is still very useful in the respect that it provides us with a conceptual framework to think about the task that the subject has to solve. So what are the predictions? First we have to be a bit more explicit on what the experiment is going to look like.

To measure discrimination performance we used a three-interval forced-choice (3-IFC) oddity task. Subjects sequentially see and feel three objects (little squares) from the two-dimensional visual-haptic stimulus space, two of which are identical in both properties and one is different in some respect. Their task is to identify the interval containing the odd stimulus. We measure discrimination performance along two axes of the two-dimensional stimulus space before and after an extensive training phase. One axis is termed the congruent axis which is the one from which stimuli in the training phase are exclusively drawn. In a pre-experiment we determined subjects' just-noticeable differences (JND) along the haptic and the visual dimension and used this information to normalize the two-dimensional stimulus space. The congruent axis is then chosen to be one of the main diagonals in this space such that for each stimulus on this line the visual and the haptic discrimination performance are the same. This also assures that we can compare the results over different subjects. The other axis we termed incongruent axis and it is perpendicular to the congruent axis. Of course, it would have been nicer to measure performance over the whole two-dimensional space but it is very cumbersome to do so.

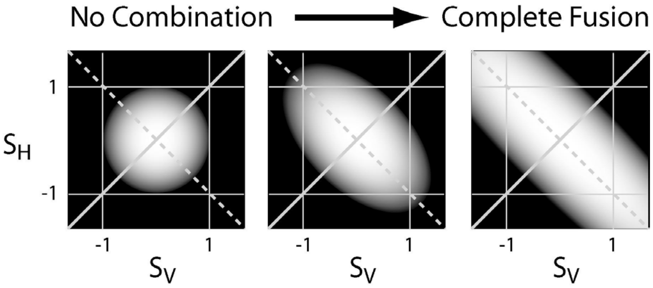

Fig. 2: Hypothetical psychometric functions for the oddity task using the MAP estimator. Three examples with different priors are shown. White corresponds to chance level and black to no error. The diagonal lines represent the congruent (solid) and incongruent (dashed) stimuli that we presented in the experiment. The variance of the prior distribution along the congruent axis is called σ_1^2, the variance along the incongruent axis σ_2^2. During the training session only stimuli from the solid line are shown. Before the training phase the subject should not assume anything about the stimuli and hence should not combine the two signals. This situation is depicted in the left panel where $\sigma_1^2 \to \infty$ and $\sigma_2^2 \to \infty$ are taken for an ignorant prior. Note that the psychometric function is circular rather than square. In the right panel it is assumed that the subject has perfectly learned that stimuli only come from the congruent line and hence his/her prior is expressed by the fact that the variance along the incongruent axis is zero: $\sigma_1^2 \to \infty$, $\sigma_2^2 = 0$. In the perfect fusion case subjects always project the signals onto the congruent axis and therefore cannot discriminate along the perpendicular direction. The middle panel depicts a more realistic case in between where $\sigma_1^2 \to \infty$ and $\sigma_2^2 = 1$. Thus, we hope to see equal thresholds along the congruent and incongruent axes before training and increased thresholds along the incongruent axis after training.

Now, choosing a decision rule we can simulate a virtual subject that uses the MAP estimates in our experiment. Given the three MAP estimates from the three intervals on what interval should the subject bet? We used the well known triangular rule as a decision rule for the oddity task: The estimate that is furthest away (in the Euclidian sense) from the center of the other two is guessed to be the odd stimulus [13]. Fig. 2 shows a simulation using this rule.

In the pre-test subjects should assume that stiffness and luminance are uncorrelated. For the purpose of illustration we can even make a stronger assumption and say that they do not know anything about the stimuli. If we measure discrimination performance along the congruent and the incongruent axes there should be no difference in discrimination performance. More formally, if we call the variance of the prior distribution along the congruent axis σ_1^2, and the variance along the incongruent axis σ_2^2 then a completely ignorant prior can be expressed by $\sigma_1^2 \to \infty$ and $\sigma_2^2 \to \infty$. This situation is depicted in the left panel of Fig. 2. In the training phase stimuli come only from the congruent axis. The hypothesis is that subjects learn that the variance σ_2^2 along the incongruent axis is reduced compared to stimuli without correlation (before training). In the extreme the subject could even learn that there is no variance along the incongruent axis whatsoever. This situation is depicted in the right panel of Fig. 2. In this case subjects always project the sensory input onto the congruent axis and cannot discriminate along the incongruent axis at all. More realistically, in the post-test we expect to see an asymmetry between the discrimination performance along the congruent and the incongruent axis that was not there before training.

In summary, we predict to find an interaction between the factors pre/post-test and congruent/incongruent if subjects can learn to combine arbitrary signals.

2 Methods

2.1 Setup

To generate the visual and haptic stimuli we used a mirror setup as depicted in Fig. 3. Subjects look onto a mirror and see a visual scene that is generated on a computer screen. Below the mirror a subject's index finger is attached to a robot arm with six degrees of freedom and force feedback along the three translatory directions (PHANToM 1.5, SensAble Technologies). Subjects have a convincing impression that they are haptically exploring the same scene they are seeing.

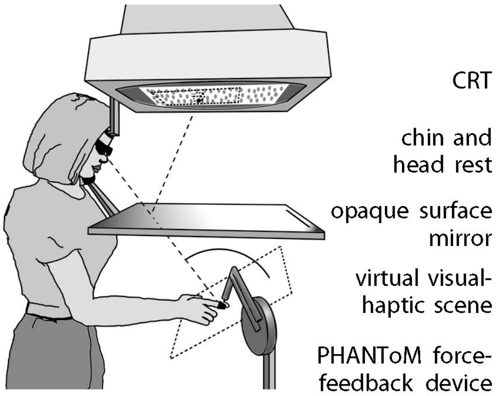

Fig. 3. The setup used can display visual scenes on a cathode ray tube (CRT), which are mirrored in order to be aligned with the haptic scene. Both scenes can be controlled independently. Haptically the scene can be explored using a PHANToM device to provide the appropriate force feedback. Subjects head is fixed on a head and chin rest. We used an SGI, Octane 2 to drive the visual and haptic simulation. GHoST was used to generate the haptic scene, OpenGL with GLUT for the visual rendering.

2.2 Stimuli and Task

Stimuli are flat squares (25mm x 25mm viewed at a distance of approximately 50 cm) that can have haptic and visual properties, namely a certain stiffness and luminance. All other properties are kept constant throughout the experiment.

- The stiffness of the square is modeled using a linear spring model with spring constant k (GHoST, SensAble, Inc.). The maximum stiffness which can be reliably generated with this device is k=0.65 N/mm. The range is normalized from 0 to 1 so the maximum k=0.65 corresponds to 1.
- For the luminance we only used the green electron beam (Sony Trinitron F500R). The exponent of the gamma correction was pre-determined with a photometer (Minolta). We were able to present 1024 different shades of green. We normalized the range from 0 to 1 so the maximum luminance 58 cd/m^2 corresponds to 1.

On a horizontal plane three flat squares are presented subsequently for 500 ms each. The subject is told that two of them are identical (visually and haptically) and one is somehow different. The subject has to guess the interval with the odd stimulus. The presentation is as follows: A white outline of the first square appears randomly on one of 16 possible locations. Once the subject touches the square he/she for 500 ms gets a sensation of stiffness and/or the square lights up with a certain luminance. 250 ms later the outline of a second square appears at another location, which the subject can see and/or feel again. The same procedure followed for a third square. After presentation the subjects made their choice.

2.3 Subjects and Groups

Twelve highly trained subjects participated (7 male, 5 female, 26.1±3.1 years, normal or corrected to normal vision, all except C.R., M.E. and F.J. naive to the purpose of the experiment). Subjects were randomly assigned into two groups, one of which got stimuli with a positive correlation between luminance and stiffness during training (hard equals bright), the other group was trained with a correlation in the other direction (hard equals dark).

2.4 Experimental Design

The experiment is a two factor within subject design. Each subject performs a pre- and post-test with a training phase in-between. Thus, one factor is performance before and after training. During training subjects were exposed to correlated stimuli only. In pre- and post-test stimuli came either from a correlated or anti-correlated distribution (two directions in visual-haptic space). That is, the second factor is the congruent or incongruent direction relative to the correlation during training. The dependent variable is the discrimination performance (threshold) in the four conditions (pre/post; congruent/incongruent).

2.5 Procedure

The experiment was divided in 4 sessions conducted on separate days. Each session lasted between 1.5 and 2.5 hours.

First Day: Normalization

In a first step we determined a subject's JND (just noticeable difference) in a purely visual and a purely haptic discrimination task. In the purely visual task the squares did not give any force-feedback and in the purely haptic task the squares were just defined by a white outline. Subjects did either the visual or the haptic task first. To measure JNDs in both visual and haptic tasks we adopted a constant stimuli procedure with a fixed standard. Each trial consisted of a fixed standard and a comparison stimulus where the odd stimulus could randomly be either the standard or the comparison stimulus.

By fitting a Gaussian to the log of the discrimination data (max. likelihood), we defined the threshold θ to be one standard deviation of this Gaussian. Besides the standard deviation we also had a nuisance parameter λ in order to account for non task-related observer lapses [14]. An example of a maximum likelihood fit (found by gradient ascent) for a purely visual task is shown in Fig. 4.

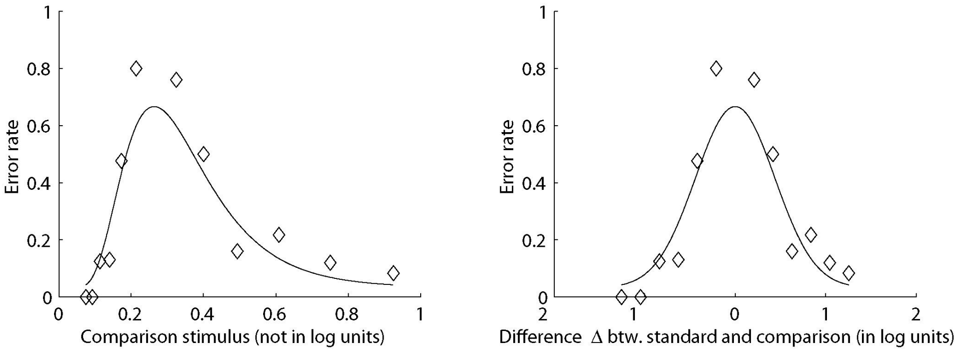

Fig. 4. Data from one subject (D.C.) in the oddity task with only haptic information available. The left panel shows error-rate vs. stiffness of the comparison stimulus. The x-axis is scaled such that the maximum stimulus intensity (in this case stiffness k=0.65 N/mm) we were able to present is set to one. The right panel shows the same data in log units with the fixed standard shifted to zero. We measured 25 repetitions per data point.

Second Day: Pre-Test

With the knowledge about individual visual and haptic thresholds for each subject we can now generate normalized stimuli individually in units of JND where we defined one JND as being the threshold that we determined on the first day. We measure discrimination performance along two directions in the normalized visual-haptic space the (+1;+1)-axis and the (+1;-1)-axis.

To measure discrimination performance in these two directions we adopt a similar procedure as described in the previous section. Again, each trial consisted of a fixed standard and a comparison stimulus where the odd stimulus could randomly be either the standard or the comparison stimulus. In order to get a measure for the discrimination performance we fit Gaussian psychometric functions. One psychometric function is used for the congruent direction and one for the incongruent direction. Thus we have one standard deviation parameter for the threshold in the congruent direction θ_c and one for the threshold in the incongruent direction θ_i. For both directions we use the same lapse rate parameter λ because data for both directions comes from the same session. Fig. 5 shows data for one subject split into congruent and incongruent trials together with a maximum likelihood fit. To get a measure of the reliability of the parameters we calculate the joint posterior of the thresholds and lapse rate $p(\theta_c,\theta_i,\lambda \,|\, n,N,\Delta)$, where n are the number of incorrect answers, N the number of trials, and Δ the difference between comparison and standard stimulus. From this distribution we integrated out the nuisance parameters we are not interested in (as priors we used a uniform distribution over a much wider range than we were expecting to see).

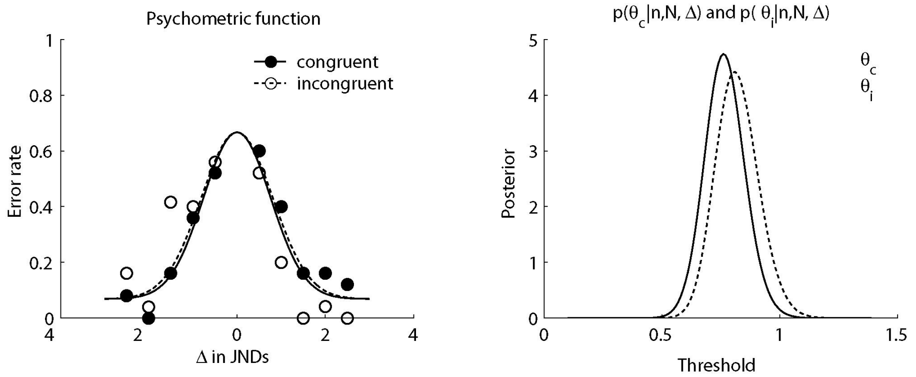

Fig. 5. Pre-Test: Discrimination performance for congruent and incongruent trials before training (subject D.C.). Distance Δ between the comparison stimulus and the fixed standard stimulus given in JND units. Left panel shows data for the congruent trials with closed circles and the maximum likelihood fit (solid line). Data from incongruent trials are depicted with open circles and the fit with a dotted line. On the right the posterior distributions for each of the threshold parameters (i.e. the joint distribution integrated over the lapse rate and the other threshold parameter) is shown. The two posterior distributions are similar and so are the thresholds for the congruent and the incongruent directions.

Third Day: Training & Post Test

During training we presented stimuli from either only the (+1;+1)-axis or from the (+1;-1)-axis depending on the group that the subject was randomly assigned to. For each group we will call the direction that is trained on the congruent direction and the other one the incongruent direction. Stimuli are equally distributed over these two directions with intensity ranging from close to zero, up to approximately the maximum we can present physically. Thus, we choose the widest possible range for the distribution of the stimuli, in order to facilitate learning of the correlation. For each trial two composite stimuli were chosen randomly from this distribution and one of them was the standard and the other one was taken as the odd stimulus. During training subjects received feedback on each trial by a beep that indicated incorrect answers. Each subject did 500 trials during this training session. It usually took subjects about an hour to complete the training session. After that subjects had a brief break before they continued with the post-test.

The post-test was identical to the procedure of the pre-test with one exception. As it is expected that during the post-test–where also half the stimuli come from the incongruent distribution–subjects will slowly unlearn what they supposedly had learned during training, we included a number of catch trials (one third of all trials) which all come from the congruent direction. Thus, there were 500 regular trials (250 congruent and 250 incongruent) plus 250 catch trials, which were all congruent to the correlation during training. Fig. 6 shows the psychometric functions for one subject determined as before.

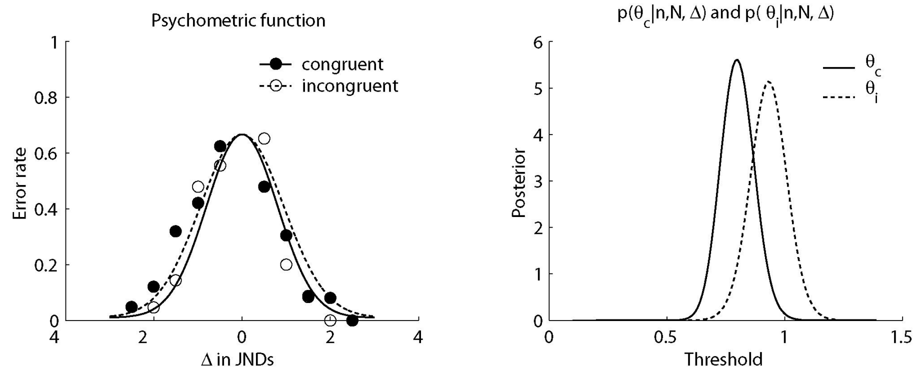

Fig. 6. Post-Test: Discrimination performance in congruent and incongruent trials after training (subject D.C.). Distance Δ between comparison and standard stimulus given in JND units. Left panel shows data for congruent trials with closed circles and maximum likelihood fit (solid line). Data from incongruent trials are represented by open circles (dotted line). The right panel shows posterior distributions over the threshold parameters of the model. Compared with pre-test distributions it is here more likely that the two parameters have different values.

Fourth day: Control

The last day was identical to the first day (Normalization). We measured performance in the task with only visual and only haptic information available to control for a general overall learning effect.

3 Results

Results for one subject are already shown in Fig. 5 and 6. For all subjects we performed the same procedure of fitting a Gaussian to the data in each direction. To get an estimate for the standard deviation and its reliability we calculated mean and variance of the posterior distributions for θ_c and θ_i in the pre-test as well as in post-test. So we had thresholds for all four conditions with the two factors: pre-test/post-test and congruent/incongruent. Three subjects (C.R., S.L.S and V.E.) were discarded from the summary analysis because their thresholds already showed large differences before training. We will discuss their results briefly in the Discussion section.

Summary results for the remaining nine subjects are depicted in Fig. 7, which shows the mean thresholds over all nine subjects for all four conditions.

An ANOVA (two factors, within subjects) on the data was conducted and revealed that there was no significant main effect, neither for pre- vs. post-test (F(1,8)= .705, p=.426) nor for congruent vs. incongruent (F(1,8)= 4.128, p=.077). However, importantly we found a significant interaction between the two factors (pre/post vs. congruent/incongruent: F(1,8)=14.58, p<.005). This shows that subjects learned to use the luminance of an object in combination with its stiffness. That is, subjects learned to combine arbitrary signals.

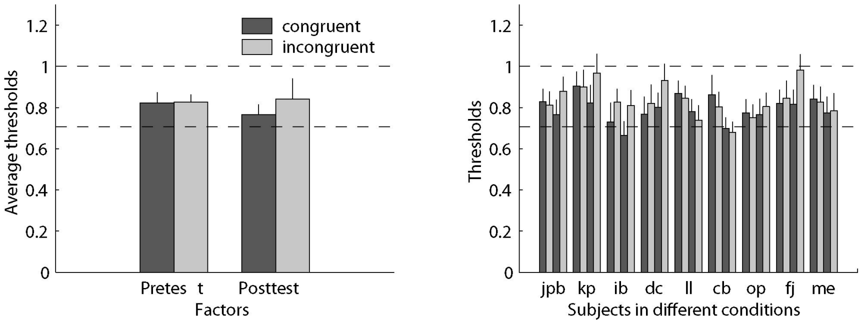

Fig. 7. Summary Results (nine subjects)**:** On the left mean and standard deviation of the thresholds across all nine subjects in all four conditions (pre/post, congruent/incongruent) are shown. The right panel shows the mean and standard deviation for each individual subject. Dark bars correspond to the congruent direction, light bars to incongruent. The left pairs of bars correspond to pre-test, the right pairs to post-test. The dotted lines show sensible upper and lower bounds for the performance. If the threshold is 1 then subjects perform as well as they would by just using one of the cues. The lower dotted line is at $1/\sqrt{2}$ (circle with radius 1 JND measured along diagonals) the best performance that theoretically can be achieved.

We also checked whether there was a significant change of thresholds over the time course of the experiment (between Day 1 and Day 4). It could be that subjects become much better in discriminating stimuli simply because they have done more than 2500 trials. We can compare the discrimination performance in the purely visual task and the purely haptic task on the first day with the performance on the last day. An ANOVA shows no significant effect for the purely visual task or the purely haptic task.

At the end of the experiment we informally queried subjects and there was no naive subject who has reported noticing the correlation during training.

4 Discussion

We conducted this study to test whether subjects can learn to combine arbitrary signals from vision and touch–namely, luminance and stiffness of an object. Therefore, we measured discrimination performance for these two signals presented simultaneously, and explored whether there is a change in discrimination performance before and after extensive exposure to a world in which these two signals are highly correlated. We predicted that the thresholds should be symmetric before training, so there is "no fusion" between these two signals. Furthermore we predicted that subjects' thresholds should become asymmetric after training if they were sensitive to the correlations in the stimuli. That is, the signals should become "somewhat fused". In case subjects' priors were fully adapted, one could say that the amount of correlation in the stimuli determines the "degree of fusion" [5].

First thing to note is that the results look reasonable. In the pre-test all subjects show a threshold below 1 JND. If a subject only used the visual or only the haptic signal for discrimination the threshold should lie at 1 JND. So at least we can say that subjects used both signals for the discrimination task. However they do not use it ideally. If they used all the available information then their threshold should lie at $1/\sqrt{2}$. If the noise is radially symmetric in the normalized visual-haptic space then the discrimination thresholds should all lie on a circle in this space.

In agreement with our predictions, we found a significant asymmetry in the discrimination performance after training, which suggests that subjects indeed learned to combine the two arbitrarily chosen signals–luminance and stiffness. The asymmetry between congruent and incongruent thresholds cannot be explained by improvement of performance due to more practice because this would have affected the congruent and incongruent direction equally. It is probably worth noting that this change in thresholds should also reflect a change of the observers' percept. This again demonstrated the high plasticity of even an adults' brain. No further significant effects were found or predicted.

Although subjects qualitatively show an asymmetry after training as we predicted based on the MAP model, the MAP model does a poor job in predicting the exact thresholds even for the average data. First of all, it has to be pointed out that the MAP model was not intended to be a model for the performance of each subject. It has to be understood as an ideal observer analysis of the cue combination problem in general. The MAP model is a normative model that shows us how the problem of combining correlated stimuli should be accomplished ideally. Obviously, subjects do not live up to these expectations. Even before training, in the uncorrelated case, they do not use all the available information.

It also has to be noted that the MAP model as we presented it should provide us with a conceptual framework rather than being the exact ideal observer for the concrete task that subjects had to solve in this experiment. The assumptions in the model do not match the experimental conditions as carefully as it they ought to if we were to attempt a proper comparison between the model and the performance of the subjects. For example, stimuli did not come from a 2d normal distribution, they were chosen as constant stimuli along only two axes. Also, during the post-test subjects un-learn what we are trying to measure with the post-test because the post-test necessarily contains incongruent stimuli.

The observed effect is small. However, we did not expect the effect to be large. Considering only one hour of training compared to a whole lifetime of experience with objects, for which luminance and stiffness are not correlated. If the system would adapt more quickly serious problems could arise. For example, if by accident stimuli get fused this had the consequence that they were harder to discriminate from one another (this is the cost of fusion and the reason why the discrimination ellipse becomes wider in the anti-correlated direction).

Three of the twelve subjects were excluded from the summary results because they showed significant differences between congruent and incongruent trials already before training. This means they already had a predefined axis for discrimination. According to the MAP model this could be due to correlated noise in the two channels, which is unlikely because luminance and stiffness are sensed by two entirely separate sensory systems (vision and touch), or it could be due to correlated priors in these subjects. Where these correlations should come from, one can only speculate. However, the important fact for our experiment is that also these three subjects produced a similar effect as the other subjects (data not shown). The difference between congruent and incongruent trials also changed in the predicted way after learning: The difference got less or disappeared.

How specific are the learned priors? For example, does the association between luminance and stiffness generalize from green squares to red circles? We don't know the answer to these questions, but it should certainly be context dependent how easy it is to change the prior. For example, it would probably be harder to change the prior for the "light from above left" [12] in comparison to the prior used in this study.

In any case, we are convinced that this framework can explain why some signals are fused whereas others are not. Objects that look big also feel big, and objects that feel small also look small. We have grown up with this sort of statistics. It would be stupid of our sensory system not to use this information.

Acknowledgement: This work was supported by the 5th Framework IST Program of the EU: Touch-HapSys (IST-2001-38040) and the Max-Planck Society.

References

1. Bülthoff, H.H., Mallot, H.A.: Integration of depth modules: stereo and shading. J. Opt. Soc. Am. A 5 (1988) 1749-1758
2. Clark, J.J., Yuille, A.L.: Data fusion for sensory information processing systems. Kluwer (1990)
3. Ernst, M.O., Banks, M.S.: Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion. Nature 415 (2002) 429-433
4. Ferguson, T.S.: Mathematical statistics – A Decision Theoretic Approach. Probability and Mathematical Statistics. Academic Press, New York (1967)
5. Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining Sensory Information: Mandatory Fusion Within, but Not Between, Senses. Science 298 (2002) 1627-1630
6. Howard, I.P., Rogers, B.J.: Interactions between depth cues. In: Howard, I.P., Rogers, B.J.: Seeing in Depth Vol. 2. I. Porteous (2002)
7. Jacobs, R.A.: What determines visual cue reliability. Trends in Cognitive Sciences 6 (2002) 345-350
8. Kersten D., Schrater P.R.: Pattern inference theory: A probabilistic approach to vision. In: Mausfeld, R., Heyer, D.: Perception and the Physical World. John Wiley & Sons, Ltd., Chichester (2002)
9. Landy, M.S, Kojima, H.: Ideal Cue Combination for Localizing Texture Defined Edges. J. Opt. Soc. Am. A 18 (2001) 2307-2320

10. Landy, M.S., Maloney, L.T., Johnston, E.B., Young, M.: Measurement and Modeling of Depth Cue Combination: in Defense of Weak Fusion. Vision Res. 35 (1995) 389-412
11. Maloney, L.T.: Illuminant estimation as cue combination. Journal of Vision 2 (2002) 493-504
12. Mamassian, P., Landy, M., Maloney, L.T.: Bayesian Modelling of Visual Perception. In: Rao, R.P.N., Olshausen, B.A., Lewicki, M.S.: Probabilistic Models of the Brain. MIT Press (2002)
13. Versfeld. N.J., Dai, H., Green, D.M.: Optimum Decision Rules for the Oddity Task. Perception & Psychophysics 58 (1996) 10-21
14. Wichmann, F.A., Hill, N.J.: The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics 63 (2001) 1293-1313
15. Yuille, A.L., Bülthoff, H.H.: Bayesian Decision Theory and Psychophysics. In: Knill, D.C., Richards, W.: Perception as Bayesian Inference. Cambridge University Press, New York (1996)