The combination of vision and touch depends on spatial proximity

Sergei Gepshtein
Vision Science Program, School of Optometry, University of California, Berkeley, CA, USA, & Laboratory for Perceptual Dynamics, Brain Science Institute, RIKEN, Japan

Johannes Burge
Vision Science Program, School of Optometry, University of California, Berkeley, CA, USA

Marc O. Ernst
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Martin S. Banks
Vision Science Program, School of Optometry and Helen Wills Neuroscience Institute, Department of Psychology, University of California, Berkeley, CA, USA

The nervous system often combines visual and haptic information about object properties such that the combined estimate is more precise than with vision or haptics alone. We examined how the system determines when to combine the signals. Presumably, signals should not be combined when they come from different objects. The likelihood that signals come from different objects is highly correlated with the spatial separation between the signals, so we asked how the spatial separation between visual and haptic signals affects their combination. To do this, we first created conditions for each observer in which the effect of combination—the increase in discrimination precision with two modalities relative to performance with one modality—should be maximal. Then under these conditions, we presented visual and haptic stimuli separated by different spatial distances and compared human performance with predictions of a model that combined signals optimally. We found that discrimination precision was essentially optimal when the signals came from the same location, and that discrimination precision was poorer when the signals came from different locations. Thus, the mechanism of visual–haptic combination is specialized for signals that coincide in space.

Keywords: haptics, inter-sensory integration, multidimensional classification, objects, optimality, proximity principle, spatial attention, vision

Introduction

The nervous system often combines information from different senses in a way that approaches statistical optimality. As a consequence, the precision of the combined estimate is better than the precision that could be derived from either sense alone (Alais & Burr, 2004; Ernst & Banks, 2002; Gepshtein & Banks, 2003; van Beers, Wolpert, & Haggard, 2002). In combining single-modality estimates, the nervous system gives more weight to the less variable estimate. Thus, a modality that affords the more precise estimate at the moment contributes more to perception than the other modalities do. In other words, the combined estimate is closer to the more precise single-modality estimate. By putting more weight on the less variable sensory estimate, the nervous system takes advantage of the fact that the precision of estimates from different modalities varies differently as a function of stimulation conditions.

However, signals from different senses should not be combined indiscriminately. Consider, for example, a person looking at one object while touching another. It is inappropriate to combine visual and haptic information in this situation because the information comes from different objects. How does the nervous system determine when to combine information from different senses in order to increase perceptual precision, and when not to combine in order to avoid combining information from different objects? This question is related to the *binding problem,* the problem of establishing a correspondence between representations in different submodalities that stem from the same object (Rosenblatt, 1961; Roskies, 1999; Treisman & Schmidt, 1982; von der Malsburg, 1999).

We investigated the inter-modality binding problem for vision and touch. We asked whether the nervous system uses the spatial proximity of visual and haptic signals to determine when they should be combined. Previous work

doi: 10.1167/5.11.7 Received June 9, 2005; published December 28, 2005 ISSN 1534-7362 © ARVO

used visual and haptic stimuli that coincided in space and found nearly optimal combination, indicated by the higher precision of the inter-modality relative to within-modality estimates (Ernst & Banks, 2002; Gepshtein & Banks, 2003). The improvement in precision is the "footprint" of combination; we used this footprint to determine when combination occurs for signals varying in their relative spatial positions.

We presented visual and haptic stimuli separated by different distances. Observers compared the sizes of two such inter-modality stimuli. If observers combined the visual and haptic signals, their performance should improve relative to their within-modality performance. We compared human performance with the performance of a model that combines single-modality signals optimally. We found that human performance approached statistical optimality when the visual and haptic signals came from the same location, and that the combination effect gradually decreased as the spatial separation between signals increased. Indeed, with sufficiently large offsets, inter-modality discrimination performance was essentially the same as within-modality performance. These findings support the view that inter-modal combination of sensory signals is specialized for object perception.

Optimal conditions for combination

To measure the effect of spatial separation between visual and haptic signals on the combination of these signals, we needed to create situations in which the effect of combining signals would be the largest. If the within-modality signals are Gaussian distributed and their noises are independent, the variance of the combined estimate with optimal weighting of the visual and haptic estimates is

$$\sigma_{VH}^2 = \frac{\sigma_V^2 \sigma_H^2}{\sigma_V^2 + \sigma_H^2}, \tag{1}$$

where σ_V, σ_H, and σ_{VH} are the standard deviations of the visual, haptic, and combined estimates (Landy, Maloney, Johnston, & Young, 1995; Yuille & Bülthoff, 1996). We define *precision* as the inverse of the standard deviation. The smaller the standard deviation is, the higher the precision. The precision of the optimally combined estimate is always higher than or equal to the highest precision of the within-modality estimates because

$$\sigma_{VH} \leq \min\{\sigma_V, \sigma_H\}.$$

The highest possible precision of the inter-modality relative to the within-modality estimates occurs when $\sigma_V = \sigma_H$. Figure 1 illustrates this: The standard deviation of the combined estimate is plotted for different values of σ_V and σ_H.

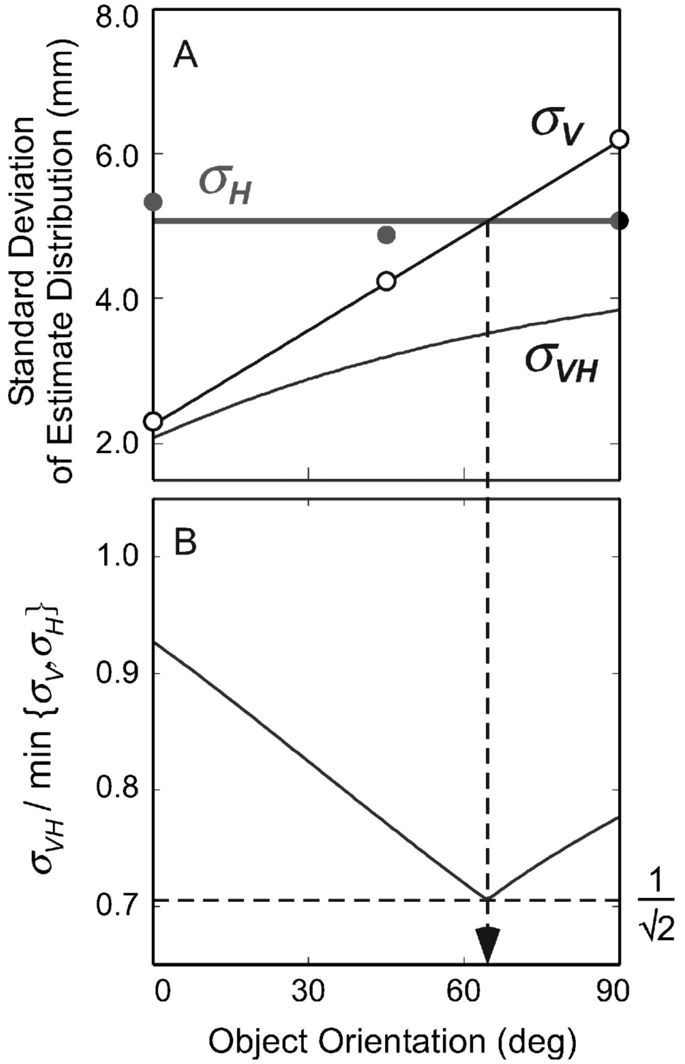

Figure 1. Precision of visual, haptic, and visual–haptic estimates as a function of object orientation. Object orientation is the slant of the parallel planes from the observer's perspective. (A) Precision of visual and haptic size estimates as a function of orientation. The gray and white dots represent the standard deviations of haptic and visual estimates, respectively (from Gepshtein & Banks, 2003). The lines are fits to those data. The curve labeled σ_{VH} is the standard deviation predicted by Equation 1; it represents the outcome of optimal visual–haptic combination. (B) The ratio of the optimal standard deviation (σ_{VH}) divided by the smaller of the within-modality deviations (σ_V or σ_H) plotted as a function of object orientation. The ratio is smallest at $1/\sqrt{2}$ when visual precision is equal to haptic precision.

Gepshtein and Banks (2003) examined whether visual–haptic estimates are optimal in the sense of Equation 1. The authors first measured size discrimination with haptics alone and vision alone, and they found that visual precision varied with object orientation while haptic precision did not (Figure 1A). The curve labeled σ_{VH} in

Figure 1A represents the inter-modality standard deviation predicted by the optimal model (Equation 1) from the within-modality measurements of Gepshtein and Banks. The ratio of the predicted standard deviation and the smallest within-modality standard deviation (visual or haptic; Figure 1B) is a measure of the expected improvement in the precision of the combined estimate relative to the within-modality estimates. When $\sigma_V = \sigma_H$, the ratio is $1/\sqrt{2}$, which is the largest possible improvement. Thus, at the object orientation for which $\sigma_V = \sigma_H$, the precision of size estimation by an observer using all the available information is better by ~29% than using only one or the other modality.

Methods

Apparatus and stimuli

The apparatus is described in Ernst and Banks (2002) and Gepshtein and Banks (2003). Visual and haptic stimuli were two planes that could be presented at different slants, but which were always parallel to one another.

The head was stabilized with a chin-and-forehead rest. Observers viewed two surfaces with both eyes and/or grasped them with the index finger and thumb to estimate the inter-surface distance. Stimulus distance from the eyes varied randomly (49–61 cm) to make the distance to one surface an unreliable cue to inter-surface distance.

The visual stimuli were random-element stereograms of two parallel planes. The simulated surfaces were 50 × 50 mm on average and were textured with uniformly distributed random dots (average radius = 2 mm, covering on average 5% of the surfaces). They were otherwise transparent. Surface area was randomized so projected area and side overlap were not useful cues to inter-surface distance. Element size and density were also randomized for the same reason. Textures were regenerated for each presentation. CrystalEyes™ liquid-crystal shutter glasses were used to present different images to the two eyes. Refresh rate was 96 Hz (48 Hz for each eye).

The haptic stimuli were generated using PHANToM™ force-feedback devices, one for the index finger and one for the thumb. The digits were attached to the corresponding PHANToM devices with a thimble and elastic band. Observers knew that the thimbles and bands were present (because we had to fit them to each digit at the beginning of an experimental run), but they quickly became unaware of them during an experimental run.

Each PHANToM device measures the 3-D positions of the tip of a digit and applies force to the digit to simulate the haptic experience of 3-D objects. In our experiments, the two PHANToM devices simulated two vertically separated planes by applying forces, normal to the planes, to the two digits. The upper simulated plane was contacted by the index finger from above, so the force was delivered upward to that digit. The lower plane was contacted by the thumb from below so that force was delivered downward. The observer's hand was not visible. Before, but not during, stimulus presentation, the tips of the finger and thumb were represented visually by small cursors; the cursors were not predictive of the inter-surface distance in the stimulus.

The haptically and visually specified separations between the planes generally differed, but the haptic planes were of the same size and orientation as the visual planes. Observers touched the haptic stimulus (the index finger from above and the thumb from below) near the horizontal midlines of the planes. They nearly always kept their digits in one position after making contact.

Observers

The same six observers with normal or corrected-to-normal vision participated in all experiments. Two (authors JDB and SSG) were aware of the experimental purpose.

Procedure

Before each trial, the observer saw two 'starter' spheres whose positions indicated the orientation of, but not the distance between, the surfaces in the upcoming trial. The observer inserted the finger and thumb into the spheres (which could be seen but not felt) and the spheres and cursors (representing the finger tips) disappeared. The disappearance was a signal to draw the finger and thumb together. In haptics-alone conditions, the observer felt two parallel (invisible) surfaces. The surfaces were extinguished 1 s after both fingers made contact. In vision-alone conditions, the movement of the fingers made both surfaces visible for 1 s (no useful haptic cue was available). In visual–haptic conditions, the observer felt and saw the surfaces simultaneously for 1 s. After the first stimulus disappeared, the 'starter' spheres reappeared, the observer inserted the fingers, and the second presentation occurred.

Two stimuli were presented on each trial: a *standard* stimulus and a *variable-size* stimulus. The standard's size was always 50 mm. The temporal order of the two stimuli was random. After the two presentations, observers indicated the one with the apparently greater inter-surface distance. No feedback was given. The visual, haptic, and visual–haptic conditions were presented in separate blocks of trials.

Before beginning the actual experiment, observers practiced the task in separate vision-only, haptics-only, and visual–haptic conditions. The practice sessions were identical to experimental sessions except that they contained only five trials per condition.

Results

Experiment 1: Finding the best orientation for each observer

In this within-modality experiment, we determined for each observer the stimulus orientation for which $\sigma_V \approx \sigma_H$. Two stimuli were presented in the center of the work space in random temporal order on each trial: a standard stimulus whose inter-surface distance was always 50 mm, and a variable-size stimulus whose inter-surface distance was 41, 44, 47, 49, 51, 53, 56, or 59 mm. Observers made a forced-choice response indicating which of the two stimuli contained the larger inter-surface distance. The value of the independent variable—inter-surface distance—was varied according to the method of constant stimuli. Each pairing of the standard and variable-size stimuli was presented 30 times to each observer.

The stimulus orientations can be expressed as surface slants relative to the line of sight. Those slants were 0, 22.5, 45, 67.5, and 90 deg. The surfaces were rotated about a horizontal axis, so the tilt (Stevens, 1983) was always 90 deg. In Experiments 2 and 3 (which were the main experiment and a control experiment, respectively), we kept the slant constant at the value determined for every observer in this experiment.

The results for one observer are shown in Figure 2A. Each panel shows the proportion of trials in which the variable-size stimulus was judged as larger than the standard as a function of the size of the variable stimulus. The top and bottom rows show data for vision only and haptics only, respectively. Each column corresponds to a different object orientation. The curves are the cumulative Gaussian functions (the psychometric functions) that best fit the data using a maximum-likelihood fitting procedure. The slope of each curve is proportional to the standard deviation (σ) of the underlying Gaussian distribution. We used σ to quantify the observer's performance in this task: The steeper psychometric function, the smaller the standard deviation of the underlying distribution and the better the discrimination. The data in the upper row of Figure 2A show that visual discrimination worsened as the object was rotated from 0 to 90 deg relative to the line of sight. The data in the lower row show that haptic discrimination did not change with orientation. Data from all observers exhibited a similar pattern (see also Gepshtein & Banks, 2003).

Figure 2B plots standard deviation of each psychometric function in Figure 2A as a function of object orientation. We will refer to the standard deviations as *just-noticeable differences,* or JNDs. We interpolated the JNDs using linear regression to find the orientation at which the visual and haptic JNDs were approximately equal (vertical arrow). As we said, testing at that orientation maximizes the expected improvement in the precision of the combined estimate relative to the within-modality estimates. We used that orientation for each observer in the subsequent experiments.

Experiment 2: Comparing inter- and within-modal performance

In the main experiment, we measured size discrimination JNDs for visual–haptic stimuli as a function of the spatial offset between the visual and haptic parts of the stimulus. The standard and variable-size stimuli were presented in random temporal order on each trial. The visual and haptic inter-surface distances (Figure 3) in the standard stimuli were always 50 mm. The visual and haptic inter-surface distances in the variable-size stimuli were equal to one another and ranged from {41, 41} to {59, 59} mm (eight values altogether). The inter-surface distance was varied according to the method of constant stimuli.

In each stimulus, the visual and haptic parts were positioned symmetrically relative to the center of the workspace. The distances from the center of the workspace to the middle of the haptic and the middle of the visual parts of stimuli were {−45, 45}, {−30, 30}, {−15, 15}, {0, 0}, {15, −15}, {30, −30}, and {45, −45} mm along the horizontal axis, yielding *spatial offsets* of −90, −60, −30, 0, 30, 60, and 90 mm. The spatial offsets were the same in the standard and variable-size stimuli presented on each trial. When the spatial offset was zero, the visual and haptics parts of the stimulus were superimposed. When the offset differed from zero, the visual and haptic parts were displaced by equal but opposite horizontal distances from the middle of the workspace (Figure 3). Thus, when the haptic part of the stimulus appeared on one side of the workspace (preceded by the visible starter spheres indicating the desired position and orientation of the hand), the observers learned to direct gaze to the corresponding position on the other side. The observers were told that the visual and haptic parts of the stimulus always came from the same object. The different offsets were presented in random order within each block of trials.

Each pairing of standard and variable-size stimuli was presented 30 times to each observer. Observers indicated which of the two stimuli contained the apparently greater inter-surface distance. No feedback was given.

Figure 4 shows the JNDs for the various conditions of the main experiment. The gray and black horizontal lines represent haptic-alone and visual-alone JNDs, respectively, from the Experiment 1, in which the stimuli were always positioned in the middle of the workspace. The dashed horizontal lines represent the JNDs predicted by the optimal combination model (Equation 1). The diamonds represent the JNDs observed with the visual–haptic stimuli. JNDs were generally smallest when the spatial offset was zero. This effect is clearest in the right panel, which plots the average JNDs for the six observers.

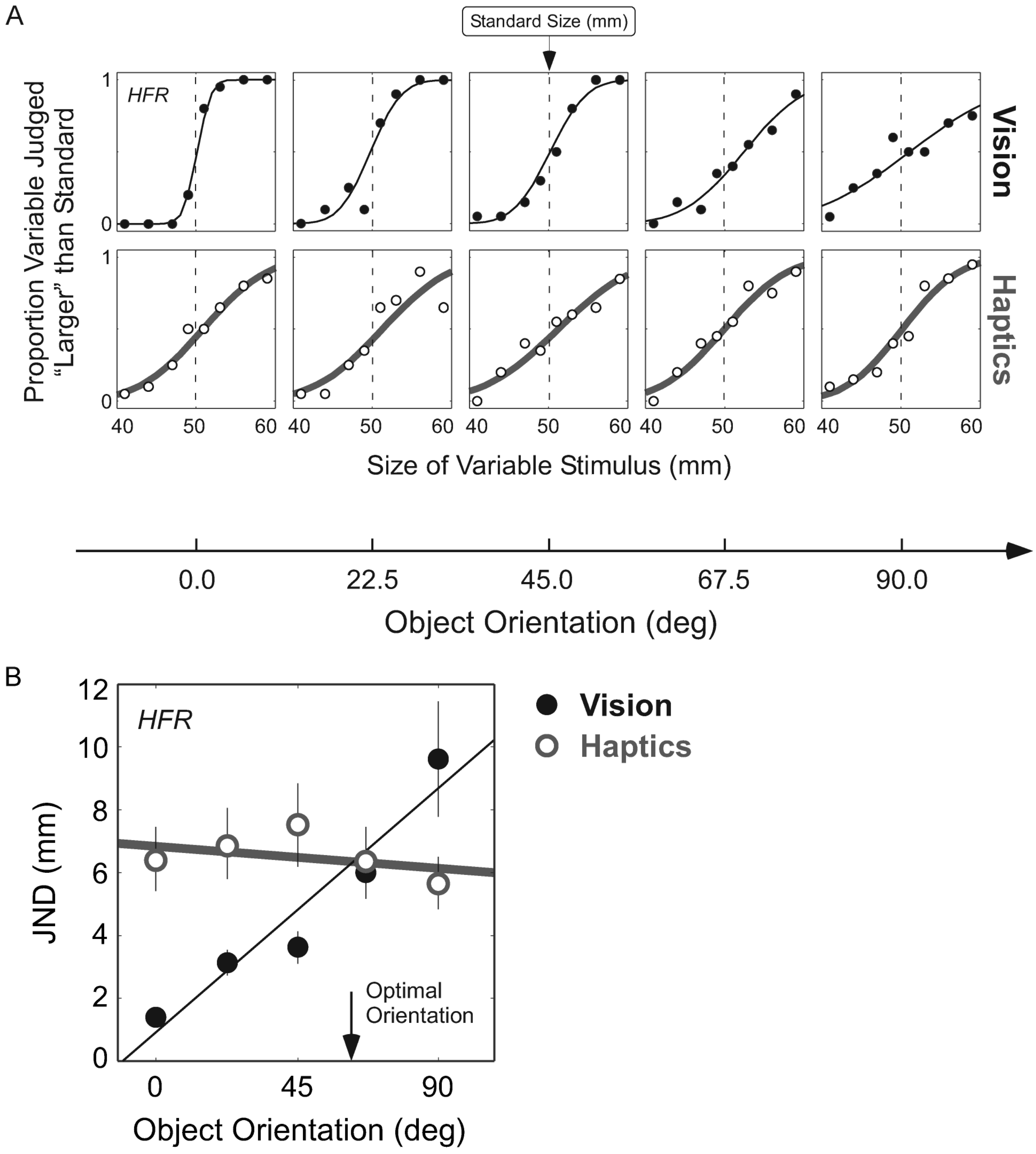

Figure 2. The results of Experiment 1 for one observer. (A) Psychometric functions for different within-modality conditions. Each panel shows the proportion of trials in which the variable-size stimulus was judged as larger than the standard stimulus as a function of the inter-surface distance of the variable-size stimulus. The top row shows data for vision only (filled symbols) and the bottom row for haptics only (unfilled symbols). Each column corresponds to a different object orientation. The curves are the cumulative Gaussian functions that best fit the data. (B) Observed visual and haptic JNDs (one standard deviation of the cumulative Gaussian functions in Panel A) as a function of object orientation. Filled circles represent the JNDs for vision alone. Unfilled circles represent the JNDs for haptics alone. We expect the precision of visual–haptic estimation to be highest at the orientation where the visual and haptic JNDs are equal, that is, where the linear regression fits to the visual and haptic data intersect. Error bars are ±1 *SE*.

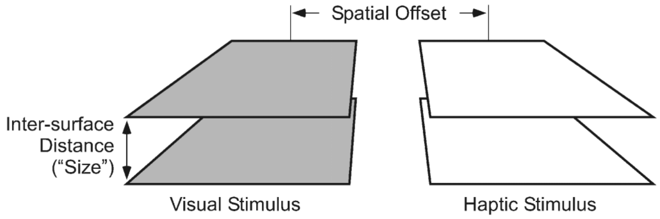

Figure 3. Schematic of the inter-modality stimulus, frontal view. The visual stimulus is on the left and the haptic on the right. The observers’ viewpoint was roughly equivalent to the viewpoint of this picture. Inter-surface distance, which observers were asked to judge, is the shortest distance between the two parallel planes; we refer to this as the stimulus “size.” Spatial offset, the main variable of interest, is the horizontal distance from the middle of the visual part to the middle of the haptic part. The visual part of the stimulus was a random-element stereogram; the parallel planes were textured with random elements. The haptic part was felt but not seen. Again the planes were parallel to one another. Stimulus orientation is the slant of the surfaces relative to the (fixed) line of sight. The object was rotated about the horizontal axis, so tilt was always 90 deg.

When the spatial offset was zero, the observed visual–haptic JNDs approached the values, one would expect for optimal combination of the visual and haptic signals. (The only exception was observer MDT, whose overall performance is better in Experiment 2 than in Experiment 1.) When the offset was large, the visual–haptic JNDs approached the within-modality JNDs. The results of statistical tests are given in the text accompanying Figures 6 and 7. The results suggest that the spatial separation between the visual and haptics parts of the stimulus helps determine whether the signals will be combined.

Experiment 3: Control for Experiment 2

There is, however, another plausible explanation for the change in JNDs with spatial offset that we observed in Experiment 2. In that experiment, we tested unimodal discrimination performance only in the center location. Perhaps the increases in JNDs at larger spatial offsets were caused by increases in the variability of the within-modality estimates at those spatial positions rather than by a breakdown in inter-modality combination. To test this possibility, we measured within-modality JNDs at three positions: −45, 0, and 45 mm from midline; these correspond respectively to the spatial offsets of −90, 0, and 90 mm in Experiment 2. This experiment was otherwise identical to Experiment 1.

The results are shown in Figure 5. The circles represent the JNDs for vision alone (filled) and haptics alone (unfilled) at the three positions. The squares represent the predictions of the optimal model at those positions for every observer (left panels) and averaged across observers (right panel). The diamonds represent the same observed visual–haptic JNDs as in Figure 4. When the spatial offset was zero, the visual–haptic JNDs were again consistently smaller than the JNDs with vision alone and with haptics alone. Presumably, the reduction of JNDs was caused by combining the two signals optimally or nearly optimally.

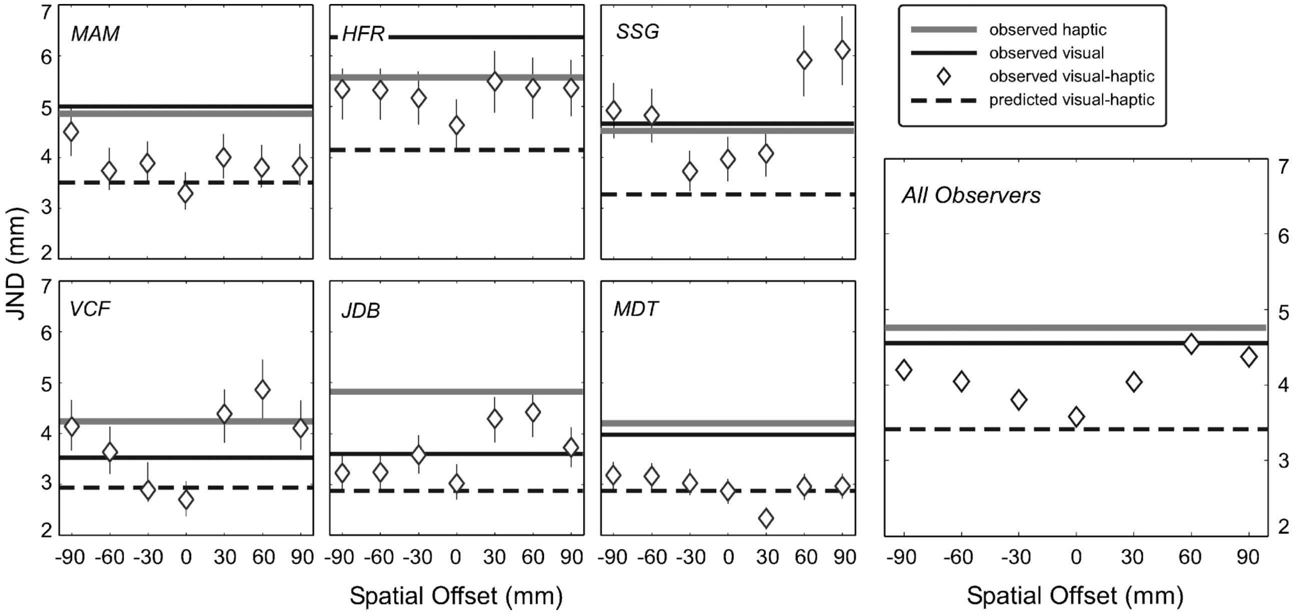

Figure 4. The results of Experiment 2: JNDs as a function of spatial offset. The six panels on the left plot JNDs for each observer. The panel on the right plots the averages across observers. The black and gray lines represent the observed JNDs for vision alone and haptics alone, respectively. The dashed lines represent the JNDs that would be predicted from the vision- and haptics-alone JNDs according to Equation 1. The diamonds are the observed visual–haptic JNDs. The error bars are ±1 *SE*.

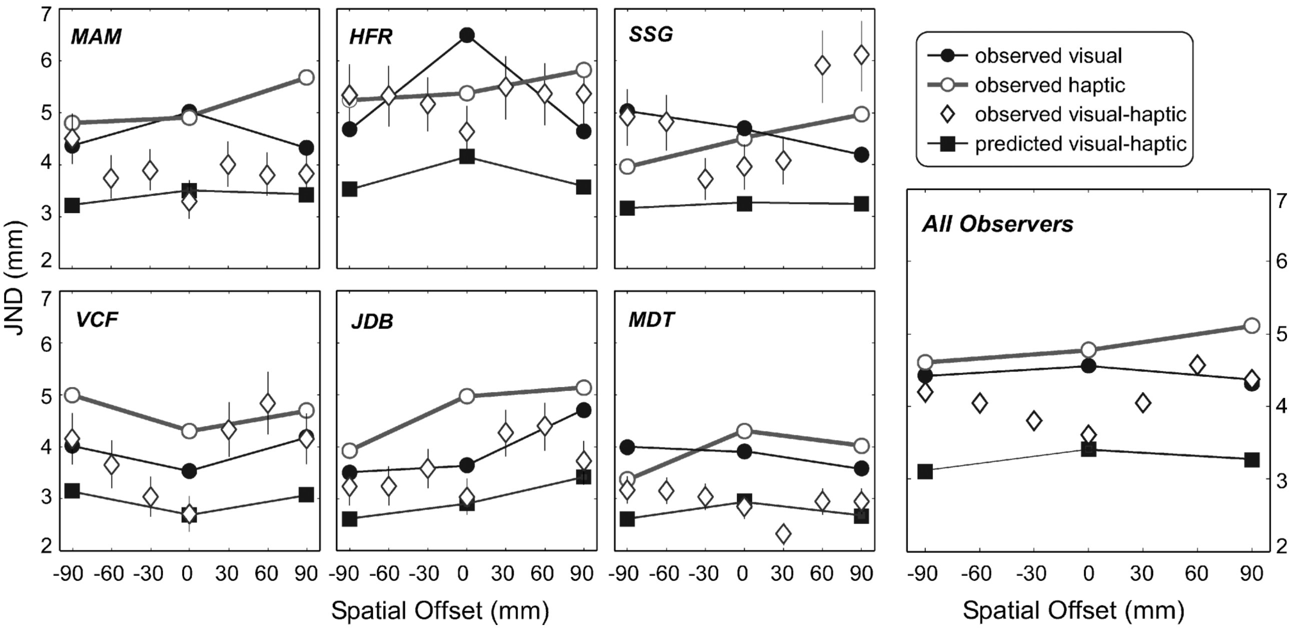

Figure 5. The results of Experiment 3: JNDs as a function of spatial offset. The diamonds are the inter-modality JNDs from Figure 4. The circles are the within-modality JNDs measured at three spatial positions of −90, 0, and 90 mm in Experiment 2. Filled circles are for vision alone and unfilled for haptics alone. The squares represent the predicted inter-modality JNDs based on the within-modality JNDs and Equation 1. As in Figure 4, the six left panels show the individual observer data and the right panel the averages across observers.

When the spatial offset was not zero, the visual–haptic JNDs approached the JNDs with vision alone and haptics alone. Presumably, that happened because the signals were not combined.

The results in Figure 5 are summarized in Figure 6. The observed within- and inter-modality JNDs and the predicted inter-modality JNDs for optimal combination are plotted as a function of the absolute value of the spatial offset. The JNDs at ±90-mm offsets were averaged for each observer to obtain the values labeled "90-mm offset". The predicted and observed inter-modality JNDs are represented by the gray and hatched bars, respectively.

The predicted and observed inter-modality JNDs were quite similar when the offset was 0 mm ($t = 0.83$, $p > .05$); additionally, the observed inter-modality JNDs were always smaller than the within-modality JNDs. The observed inter-modality JNDs were always higher than the predicted JNDs when the offset was |90| mm ($t = 8.42$, $p < .001$); they became similar to the within-modality JNDs.

Figure 7 plots the results across observers. Observed inter-modality JNDs are plotted against predicted JNDs. The diagonal line represents perfect agreement between observed and predicted JNDs. The zero-offset data are much closer to that line (reduced $\chi^2 = 0.54$) than the |90|-mm offset data (reduced $\chi^2 = 4.18$; Bevington & Robinson, 1992). Thus, observers combined visual and haptic estimates in a nearly optimal fashion when the offset was zero and did not when it was |90| mm.

Discussion

Summary of results

Size discrimination with visual–haptic stimuli was most precise when visual and haptic signals were spatially coincident. In fact, when the signals were coincident, discrimination performance was statistically indistinguishable from optimal (Equation 1). When they were not coincident, visual–haptic discrimination precision decreased: At large spatial offsets, it was as low as the precision with one sense alone. Thus, the spatial separation between visual and haptic signals is one factor that determines whether the nervous system combines visual and haptic signals.

Inter-sensory object perception

The visual system correctly interprets most images it receives from the environment in part because of the perceptual grouping mechanisms that link image features arising from the same physical source (Elder & Golsberg, 2002; Martin, Fowlkes, Tal, & Malik, 2001; Ruderman & Bialek, 1994). Features that are near one another spatially tend to come from the same object and be linked perceptually. Spatially separated features tend to come from different objects and not be linked perceptually (Geisler, Perry, Super, & Gallogly, 2001; Kubovy, Holcombe, &

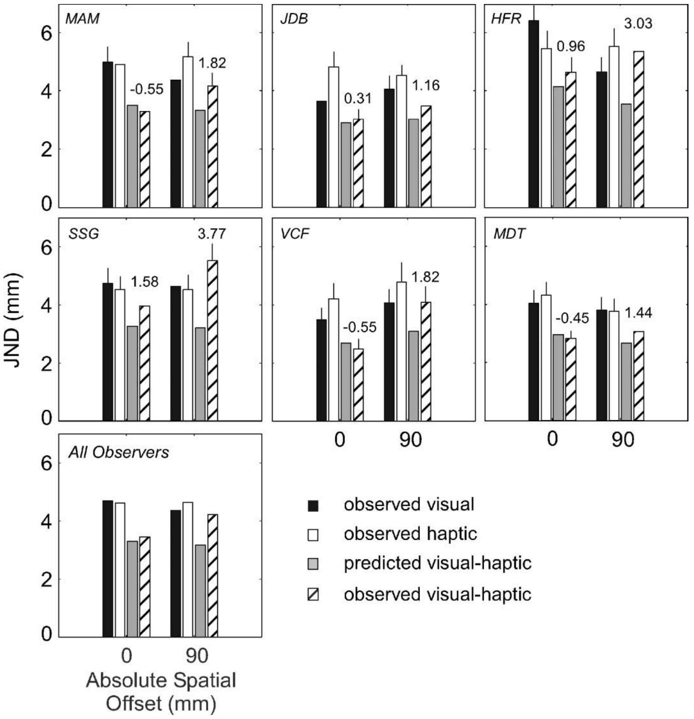

Figure 6. Observed and predicted JNDs as a function of the absolute value of the spatial offset. The upper six panels show JNDs from the individual observers and the bottom panel shows JNDs averaged across observers. The black and white bars represent the observed visual and haptic JNDs, respectively. At the offset of 0 mm, the stimuli were presented at midline. At the offset of |90| mm, they were presented |45| mm away from midline (corresponding to spatial offsets of |90| mm in the inter-modality conditions of Experiment 2). The gray and hatched bars represent the predicted and observed visual–haptic JNDs, respectively, for those positions. The numbers above the hatched bars are the difference between the observed and predicted inter-modality JNDs divided by the standard error of the estimates of the inter-modality JNDs.

Wagemans, 1998; Wertheimer, 1923). The work reported here shows that visual and haptic signals are more likely to be combined when they are spatially coincident. Thus, our results are clearly related to the visual proximity principle in perceptual organization. As with the visual proximity principle, using spatial proximity as a cue for inter-sensory combination should aid everyday object perception by maximizing the probability that signals from the same rather than different objects are combined.

The model generally used in inter-sensory cue combination (e.g., Ernst & Banks, 2002) states how sensory precision should increase when inter-sensory signals are combined. The model does not incorporate the spatial proximity of the signals. Our results suggest that a more general model is needed: a model in which the mechanism of cue combination takes into account cues (such as spatial proximity) indicating whether the inter-sensory signals come from the same object.

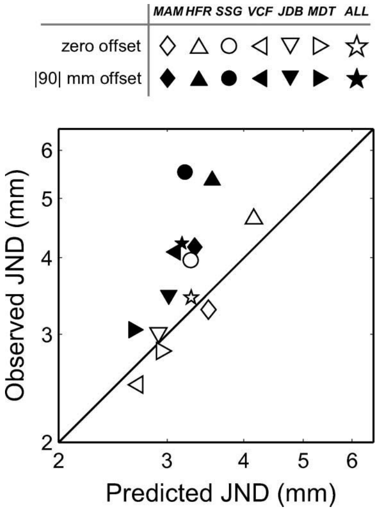

Figure 7. Observed JNDs as a function of the predicted JNDs. The symbols (except for the stars) represent the values for different observers. The stars represent the averages across observers. The diagonal line is the line of perfect agreement between the predicted and observed JNDs (see text for statistical details).

Factors influencing signal combination

There are many properties of signals that are likely to affect the nervous system's ability to combine information from different senses. In the work presented here, we showed that spatial separation between visual and haptic signals affects this ability. Gepshtein and Banks (2003) showed that the difference in size between visual and haptic signals affects the ability for visual–haptic combination as well. In that study, observers made size judgments between spatially coincident visual and haptic signals. Gepshtein and Banks varied the *conflict* between the two signals: the difference in the sizes specified by vision and haptics. Visual–haptic discrimination performance was best when the conflict was zero and became successively poorer as the conflict became larger (their Figure S2). Other studies have found that separation in time also affects the ability to combine signals (Bresciani et al., 2005; Shams, Kamitani, & Shimojo, 2000).

Taken together, the present results and those of the previous studies suggest that the nervous system determines when to combine visual and haptic signals based on signal similarity: similarity of spatial position, similarity of size, and similarity in time. Thus, to determine whether to combine signals from different modalities, the nervous system is solving a classification problem (Duda, Hart, & Stork, 2001). Because signals from different modalities vary along many dimensions, it is a multidimensional classification problem. Such a problem is often solved by computing a measure of signal similarity that takes into account signal differences on multiple dimensions (Coombs, Dawes, & Tversky, 1970; Krantz, Luce, Suppes, & Tversky, 1971). Such a measure could be used by the nervous system to determine whether to combine the signals. To further investigate how signal similarity in several dimensions affects the integration of visual and haptic information, one could examine the precision of a multi-modal estimate while varying the stimulus along several sensory dimensions, as we did here for one dimension. A satisfactory model of this process would have a measure of signal similarity that reliably predicts the precision of the multi-modal estimate. In that case, different combinations of signal parameters (e.g., visual and haptic size, location, time of occurrence, etc.) that correspond to the same similarity value should yield the same precision.

It would be interesting to know whether inter-sensory combination is affected by higher-level variables such as occlusion relationships, or whether it is affected by only low-level variables such as spatial proximity. For example, imagine that an occluder is placed in front of the gap between the visual and haptic parts of our stimulus. With amodal completion (Kanizsa, 1979), the two parts might appear to belong to the same object. Would observers then combine more widely separated visual and haptic signals than we observed? Such a finding would suggest that high-level variables are indeed involved in inter-sensory combination.

What causes the gradual effect of spatial separation?

We observed a gradual rather than abrupt change in the amount of inter-sensory combination as spatial separation was increased. The most likely cause of this gradual effect is statistical: If signal similarity was not reduced on any other dimension (e.g., temporal similarity), the signals might always be combined when the spatial offset is zero, never combined when the offset is large, and combined some of the time at intermediate offsets. If this occurred, a gradual effect of spatial separation would be observed as in our experiments.

Are the results a manifestation of spatial attention?

The inter-modality task required attending to both visual and haptic information. If we make the common assumption that attention has a limited spatial extent, then the separation of the visual and haptic signals should have affected how attention was allocated to the two signals. When the signals were in the same location, attention could be directed to one region in space. When they were in different locations, attention either had to be divided or its spatial extent had to be expanded in order to incorporate both locations. If we make the additional reasonable assumption that dividing or expanding attention leads to greater variability in sensory estimates (Prinzmetal, Amiri, Allen, & Edwards, 1998), we would predict better discrimination performance when the visual and haptic signals coincided and poorer performance when they did not.

This divided attention (or expanded attention) account does not contradict the combination model presented in Equation 1. Rather, the ability to devote attention to visual and haptic signals when the signals are coincident could be part of the mechanism by which inter-modality combination occurs. And the inability to divide attention to two different locations when the signals are not coincident could be part of the mechanism by which inter-modality combination does not occur. Along these lines, Macaluso, Frith, and Driver (2001) and Spence, McDonald, and Driver (2004) have argued that inter-modality attention and inter-modality integration are mediated by the same neural substrate.

Do the results manifest a unified multi-modal percept?

The improvement in precision observed in the inter-modality experiment could in principle result from a perceptual process or a decision strategy. By the former, we mean that the observer's judgments are based on a unified multi-modal estimate resulting from the weighted combination of visual and haptic signals (Hillis, Ernst, Banks, & Landy, 2002). By the latter, we mean that the observer's decision is based solely on comparing (and weighting appropriately) the two unimodal signals without actually combining them into a unified percept. That is, the information could still be used optimally, but without the percept of a single object. Our study cannot distinguish these two possibilities because they could both be affected by spatial proximity.

Conclusions

We examined the rules that govern the combination of signals from two different senses. When visual and haptic signals were presented in the same location, combination occurred and this yielded an improvement in perceptual precision that approached statistical optimality. When visual and haptic signals were separated by more than ~3 cm, combination did not seem to occur because perceptual precision was no better than the precision expected from vision or haptics alone. Thus, the spatial separation of visual and haptic signals is one factor that determines whether the nervous system combines signals from different senses.

Acknowledgments

This research was supported by grants from NIH (EY12851), AFOSR (F49620-01-1-0417), and Silicon Graphics. Part of this work was presented at the annual Vision Sciences Society meeting, 2004.

Commercial relationships: none.
Corresponding author: Sergei Gepshtein.
Email: sergei@brain.riken.jp.
Address: Laboratory for Perceptual Dynamics, Computational Neuroscience Group, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan.

References

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

Bevington, P., & Robinson, D. K. (1992). *Data reduction and errror analysis for the physicals sciences*. New York: McGraw-Hill.

Bresciani, J. P., Ernst M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: Auditory signals can modulate tactile taps perception. *Experimental Brain Research, 162,* 172–180. [PubMed]

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. John Wiley & Sons.

Elder, J., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision,* 2(4), 324–353, http://journalofvision.org/2/4/5/, doi:10.1167/2.4.5. [PubMed] [Article]

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433. [PubMed]

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41,* 711–724. [PubMed]

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

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between senses. *Science, 298,* 1627–1630. [PubMed]

Kanizsa, G. (1979). *Organization in vision*. New York: Praeger.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. 2*. New York: Academic Press.

Kubovy, M., Holcombe, A. O, & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology, 35,* 71–98. [PubMed]

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measuring and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35,* 389–412. [PubMed]

Macaluso, E., Frith, C., & Driver, J. (2001). A reply to McDonald, J. J., Teder-Sälejärvi, W. A., & Ward, L. M. Multisensory integration and crossmodal attention effects in the human brain. *Science, 292,* 791a.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International conference on computer vision* (pp. 416–425). Los Alamitos, CA: IEEE Computer Society Press.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). The phenomenology of attention: Part 1. Color, location, orientation, and "clarity." *Journal of Experimental Psychology. Human Perception and Performance, 24,* 261–282.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptions and the theory of brain mechanisms*. Washington, DC: Spartan Books.

Roskies, A. L. (1999). Introduction to the binding problem. *Neuron, 24,* 7–9.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters, 73,* 814–817. [PubMed]

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408,* 788. [PubMed]

Spence, C., McDonald, J., & Driver, J. (2004). Exogenous spatial-cuing studies of human crossmodal attention and multisensory integration. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 277–320). New York: Oxford University Press.

Stevens, K. A. (1983). Surface tilt (the direction of surface slant): A neglected psychophysical variable. *Perception & Psychophysics, 33,* 241–250. [PubMed]

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107–141. [PubMed]

van Beers, R. J., Wolpert, D. M., & Haggard, P. (2002). When feeling is more important than seeing in sensorimotor adaptation. *Current Biology, 12,* 834–837. [PubMed]

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron, 24,* 95–104. [PubMed]

Wertheimer, M. (1936). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of gestalt psychology* (pp. 71–88). Routledge & Kegan Paul, London. (Original work published in 1923)

Yuille, A. L., & Bütlhoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–161). Cambridge University Press.