Current Biology, Vol. 13, 483–488, March 18, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/S0960-9822(03)00133-7

Viewing Geometry Determines How Vision and Haptics Combine in Size Perception

Sergei Gepshtein* and Martin S. Banks*
University of California, Berkeley
Vision Science Program
School of Optometry
Berkeley, California 94720-2020

Summary

Vision and haptics have different limitations and advantages because they obtain information by different methods. If the brain combined information from the two senses optimally, it would rely more on the one providing more precise information for the current task. In this study, human observers judged the distance between two parallel surfaces in two within-modality experiments (vision-alone and haptics-alone) and in an intermodality experiment (vision and haptics together). In the within-modality experiments, the precision of visual estimates varied with surface orientation, as expected from geometric considerations; the precision of haptic estimates did not. An ideal observer that combines visual and haptic information weights them differently as a function of orientation. In the intermodality experiment, humans adjusted visual and haptic weights in a fashion quite similar to that of the ideal observer. As a result, combined size estimates are finer than is possible with either vision or haptics alone; indeed, they approach statistical optimality.

Results

The precision of perception varies in everyday settings. For example, changes in viewing distance, lighting, and motion affect the ability to estimate object properties visually. Consider estimating the distance between two parallel planar surfaces. When the surfaces are parallel to the line of sight (Figure 1A), visual estimation is straightforward: the retinal angle between the projections of the two surfaces is measured and scaled for distance. In this case, the error in estimating intersurface distance should increase in proportion to viewing distance. When the surfaces are perpendicular to the line of sight (and transparent; Figure 1B), visual estimation is more difficult: now one must measure binocular disparity between the surfaces and scale for distance. Because of the geometric relationship between disparity and relative distance, the error in estimating intersurface distance should increase in proportion to the square of viewing distance. Thus, we expect visual judgments of intersurface distance to be more precise in the former than in the latter case [1, 2]. If the observer estimates the intersurface distance haptically (active touch), she rotates the wrist to place the finger and thumb in the appropriate orientation. The proprioceptive and efferent signals from the digits as they contact the surfaces are similar in the parallel and perpendicular cases; so, in this situation, the precision of haptic estimates should not vary with orientation (see [3] for a counter example).

Suppose the observer looks at and feels the surfaces simultaneously. The principle of maximum likelihood (ML) prescribes the strategy for combining visual and haptic estimates that produces the estimate of lowest variance [4–8]. If the visual and haptic estimates are independent and normally distributed, that strategy is weighted summation

$$\hat{S}_{VH} = w_V\hat{S}_V + w_H\hat{S}_H,$$
$$w_v = \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_H^2}, \quad w_H = \frac{1/\sigma_H^2}{1/\sigma_V^2 + 1/\sigma_H^2}, \qquad (1)$$

where $\hat{S}_V$, $\hat{S}_H$, and $\hat{S}_{VH}$ are the visual, haptic, and combined estimates, respectively. The ws and σs are the weights and standard deviations of the estimates, respectively. According to this model, the combined estimate is shifted toward the estimate of lower variance. Thus, if the visual estimate is more precise than the haptic, the optimal combined estimate would be closer to the visual size. If the visual estimate is less precise than the haptic, the optimal combined estimate would be closer to the haptic size. The variance of the combined estimate is

$$\sigma_{VH}^2 = \frac{\sigma_V^2\sigma_H^2}{\sigma_V^2 + \sigma_H^2}, \qquad (2)$$

which is lower than the haptic and visual variances. Thus, the optimal combination is more precise than either vision or haptics alone [7, 8].

Do humans combine vision and haptics optimally? To find out, we varied the orientation of two parallel surfaces and had people judge the distance between the surfaces. We first asked whether precision varies with surface orientation when only visual information is available (as expected from the viewing geometry), and whether precision is constant across orientation when only haptic information is available (as expected from hand mechanics). Then, from the within-modality measurements (vision-alone and haptics-alone), we determined the optimal weights (Equation 1) for intermodality (visual-haptic) measurement. We then conducted a visual-haptic experiment to determine whether humans combine information across the senses in a statistically optimal fashion.

The visual stimuli were random-element stereograms of two parallel planes under three orientations—parallel, oblique, and perpendicular—relative to the line of sight. The haptic stimuli were created by using force-feedback devices, one each for the index finger and thumb. The visual and haptic stimuli were superimposed in the workspace. Observers could not see their hand.

Within-Modality Experiment

In a two-interval, forced-choice procedure, observers reported which of two 750-ms presentations contained

*Correspondence: sergeg@uclink.berkeley.edu (S.G.); marty@john.berkeley.edu (M.S.B.)

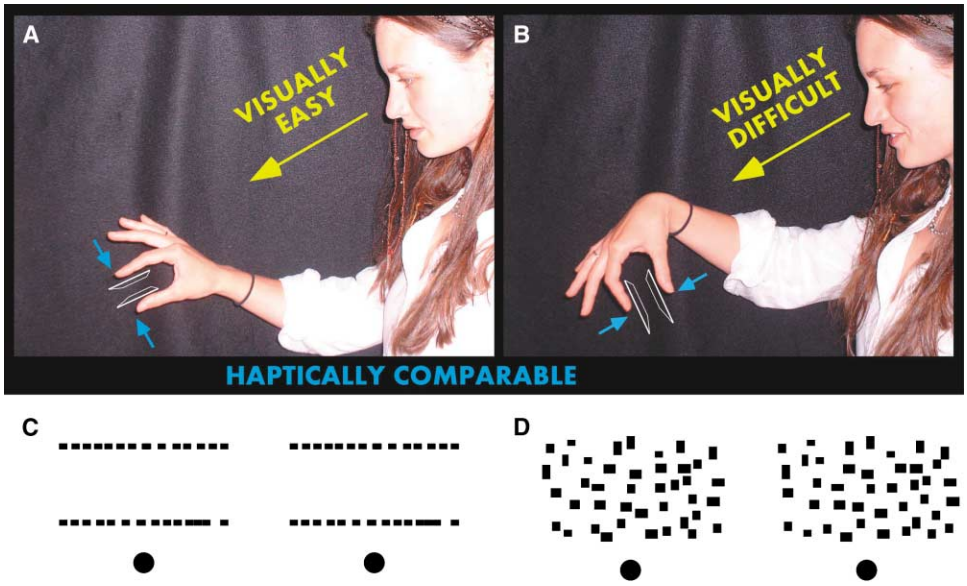

Figure 1. Estimating the Distance between Two Parallel Surfaces

(A and B) For vision, the task is presumably easier on the left (surfaces parallel to the line of sight) than on the right (perpendicular). For touch, the difficulty is presumably similar in the two cases.

(C and D) The diagrams below are stereograms depicting the visual stimuli. To view them, converge or diverge the eyes.

the stimulus with the larger intersurface distance. The size of one stimulus, the standard, was always 50 mm; the size of the other, the comparison, varied. The experiment was conducted in two blocks: vision-alone and haptics-alone. Figure 2 shows the results, averaged across observers. Figures 2A and 2B show the proportion of trials for which the comparison was judged as larger than the standard as a function of the comparison distance. The slopes of cumulative normals fitted to the data correspond to the precision of the within-modality judgments: steeper slopes indicate greater precision. As expected, precision with vision-alone was highest when the surfaces were parallel to the line of sight and lowest when they were perpendicular. Also, as expected, precision did not vary with orientation in the haptics-alone condition. The just-noticeable differences (JNDs) are plotted in Figure 2C. JNDs for vision alone increased as orientation changed from parallel to perpendicular; haptic JNDs did not change.

Intermodality Experiment

We next asked whether the brain fully utilizes visual and haptic information when both are available. Specifically, does vision receive more weight than haptics when the surfaces are parallel to the line of sight, and does haptics receive more weight than vision when the surfaces are perpendicular? We presented visual and haptic information specifying intersurface distance. To determine the weights, we introduced a discrepancy between the visually and haptically specified distances.

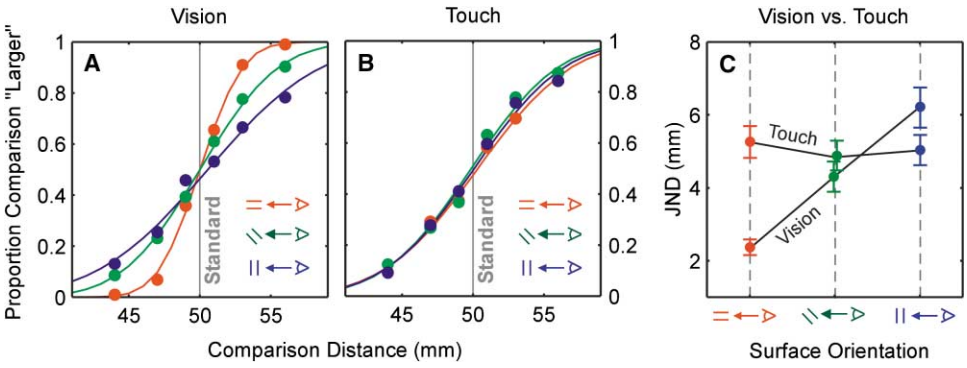

Figure 2. Results of the Within-Modality Experiment

(A and B) The proportion of trials in which the comparison was judged as larger than the standard as a function of the comparison's intersurface distance. Red, green, and blue symbols and curves correspond to data from the parallel, oblique, and perpendicular conditions, respectively. (A) and (B) show the data from the vision-alone and haptics-alone conditions, respectively. The curves are cumulative normals that best fit the data once averaged across the five observers.

(C) Observed JNDs (1 SD of the cumulative normals in [A] and [B]) as a function of surface orientation. Error bars are ± 1 SE.

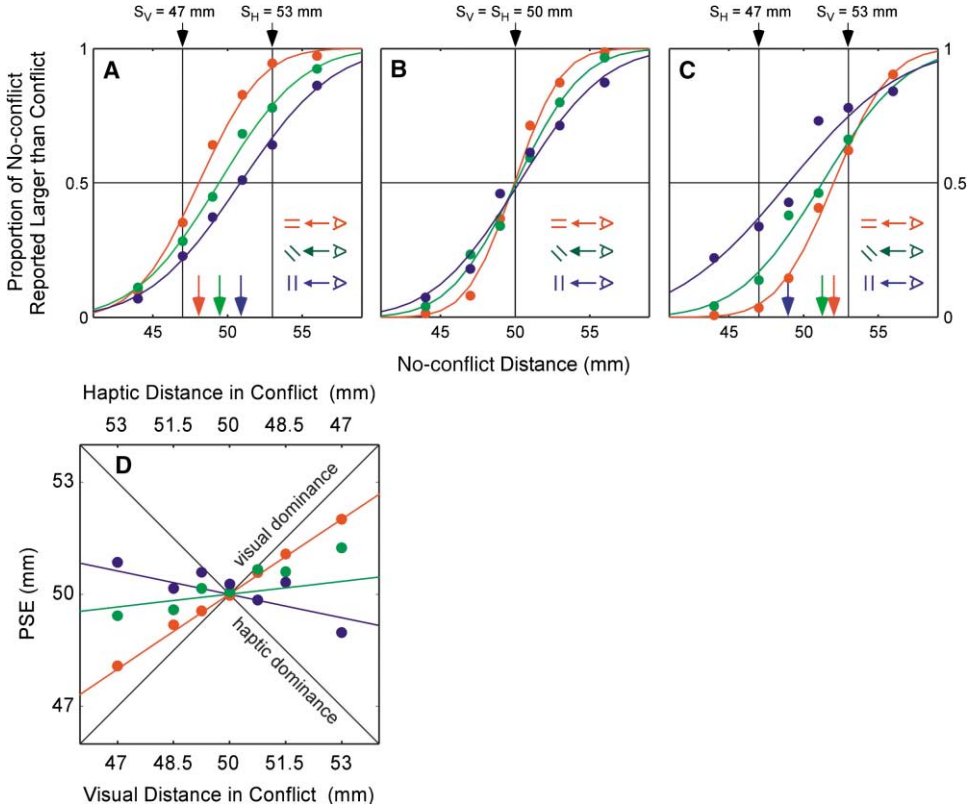

Figure 3. Results of the Intermodality Experiment

(A–C) The proportion of trials in which the no-conflict stimulus was judged as larger than the conflict stimulus is plotted as a function of the intersurface distance in the no-conflict stimulus. The data have been averaged across observers. (A), (B), and (C) show the data for conflict pairings (visual-haptic) of {47, 53}, {50, 50}, and {53, 47} mm, respectively (3 of the 7 conflicts). The red, green, and blue symbols are data from the parallel, oblique, and perpendicular conditions, respectively. The curves are cumulative normals that best fit the averaged data. PSEs are the values of the no-conflict stimulus for which the observer reports that it is larger than the conflict stimulus half the time. Those values are indicated for the parallel, oblique, and perpendicular conditions by the red, green, and blue arrows, respectively.

(D) Predicted and observed PSEs plotted as a function of the visually specified distance (lower abscissa) or haptically specified distance (upper abscissa) in the conflict stimulus. The diagonal gray lines show the predicted PSEs if vision or haptics completely dominated the combined percept. PSEs predicted by the ML combination rule (Equation 1) are represented by the colored lines (derived from the within-modality data averaged across observers). The circles represent the observed PSEs, averaged across observers. (The effect of stimulus orientation on PSEs was highly significant, $p < 0.01$, as indicated by multiple regression of the PSE data on intersurface distance and stimulus orientation, $R^2 = 0.93$. PSEs for individual observers are shown in Figures S1 and S3 in the Supplemental Data).

Two 750-ms stimuli, no-conflict ($S_V = S_H$) and conflict ($S_V \neq S_H$), were presented in random order. Observers indicated the one containing the larger intersurface distance. Figure 3 shows the results; Figures 3A–3C show the proportion of trials in which the no-conflict stimulus was judged as larger than the conflict stimulus as a function of the no-conflict size. The psychometric functions and points of subjectively equal size (PSEs) were shifted toward the visual size in the parallel condition and toward the haptic size in the perpendicular condition. These shifts are consistent with the expectation that vision will dominate the judgment when the visual variance is lower than the haptic variance and that the reverse will occur when the visual variance is higher.

We next examined how closely the visual-haptic data conformed to the predictions of ML combination. Using the visual and haptic variances (σ_V^2 and σ_H^2) measured in the within-modality experiments, we calculated the predicted PSEs (Equation 1); these are represented by the colored lines in Figure 3D. If vision completely dominated the visual-haptic percept, the visually specified distances of the conflict and no-conflict stimuli would have to be physically equal to be perceived as equal; the data would have a slope of 1. Similarly, complete haptic dominance would yield data with a slope of -1. If neither vision nor haptics completely dominated, the PSEs would fall between the diagonals. The ML prediction is closest to visual dominance when surface orientation was parallel to the line of sight because visual estimates were most precise in that condition. The prediction shifted toward haptic dominance when the stimulus was perpendicular because vision was less precise than haptics in that condition (Figure 2). The data points represent the observed PSEs. The agreement between predicted and observed PSEs is very good. The best-fitting slopes for observed and predicted PSEs are 0.65 (SE = 0.04) and 0.67 for the parallel condition, 0.31 ($\pm$0.09) and 0.115 for the oblique, and -0.25 ($\pm$0.20) and -0.21 for the perpendicular condition. The observed and predicted PSEs are statistically indistinguishable, except in the

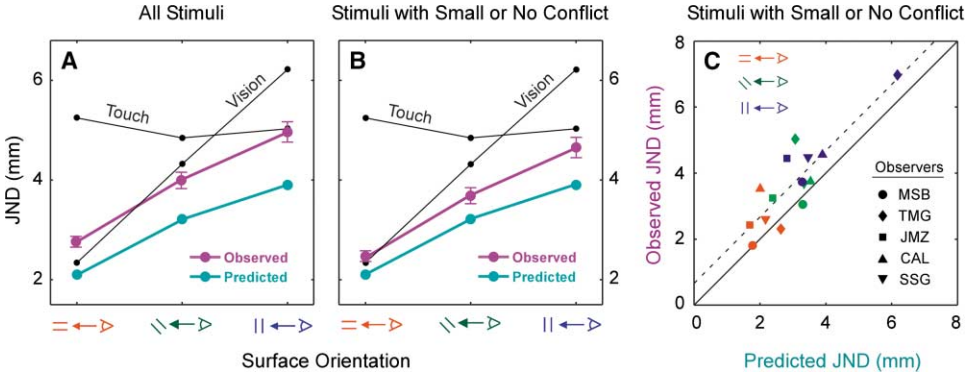

Figure 4. Precision of Distance Estimates in the Intermodality Experiment

(A and B) Observed and predicted JNDs plotted as a function of surface orientation. Black points and lines represent observed JNDs in the within-modality experiment (see also Figure 2). Cyan points and lines represent predicted JNDs, and purple points and lines represent observed JNDs in the intermodality experiment. (A) shows the averaged data from all conditions. (B) shows the averaged data from the smallest conflicts (1.5 mm or less). Error bars are ± 1 SE. The observed intermodality JNDs are significantly smaller than the smallest within-modality JNDs in the oblique and perpendicular conditions ($p < 0.05$, z scores of their differences are 2.15 and 2.07). In the parallel condition, however, the observed intermodality JND is indistinguishable from the visual JND (z score of their difference is -0.58).

(C) Observed and predicted JNDs for each observer from the smallest conflicts. The red, green, and blue symbols represent data from the parallel, oblique, and perpendicular conditions, respectively, for individual observers. The solid diagonal is the line of perfect agreement. The dashed line is a least-squares linear fit to the observed JNDs. It has an intercept of 0.66 mm and a slope of 1; it indicates that, on average, humans are 0.66 mm less precise than ideal. The correlation between the predicted and observed is 0.85; by the estimate of Pugh and Winslow [20], the probability of 15 measurements of 2 uncorrelated variables yielding such a correlation is less than 0.0005.

oblique condition, where vision was given too much weight relative to prediction. Overall, the PSEs suggest that the brain is nearly optimal statistically in taking varying visual precision into account. One cannot, however, determine from average responses (such as PSEs) whether the variability of the combined estimate is reduced relative to the vision-alone and haptics-alone estimates. To examine this, we looked at how discrimination thresholds (JNDs) were affected.

An observer following the ML combination rule would make finer discriminations when vision and haptics were both available than when only one was (Equation 2). The precision is given by the JND (slope of psychometric function, Figure 3). Figures 4A and 4B plot observed within-modality JNDs and predicted and observed intermodality JNDs. Figure 4A includes all the data from the intermodality experiment. The observed thresholds in that experiment were similar to or lower than the visual and haptic thresholds in the within-modality experiments, but the observed thresholds were not quite as low as predicted. This may have resulted from occasional awareness of the discrepancy between the visual and haptic stimuli. Perhaps such awareness caused observers to adopt a less-than-optimal strategy (such as switching between only the visual or only the haptic percept when the conflict was noticeable, [7]). To test this, we reanalyzed the data by using only trials in which the conflict was 1.5 mm or less. Figure 4B shows the result: the observed thresholds were closer to the predictions. (See Figure S2 in the Supplemental Data available with this article online for further analysis.)

Figure 4C shows the predicted and observed JNDs for small or zero conflicts for each observer and each stimulus orientation. The good agreement between predicted and observed shows that individual differences in intermodal discrimination can be largely explained by behavior in the within-modality experiments.

The finding of close correspondence between observed and predicted thresholds shows that humans combine visual and haptic information in a fashion that allows finer discrimination than is possible from either sense alone. Indeed, by the criterion of discrimination capability, the combination approaches statistical optimality.

Discussion

We have shown that the nervous system acts as if it reassigns the weights of visual and haptic estimates when the reliability of the visual estimate changes. A similar argument has been made in other domains: perception of depth, slant, and curvature from eye position and vertical disparity [9–13], depth from texture and motion [14], shape from disparity and texture [15, 16], and perception of hand position from proprioception and vision [2, 17]. These reports either did not make quantitative predictions about the combined percept or used free modeling parameters to fit the data. Only two measured the component reliabilities separately, used those measurements to generate quantitative predictions with no free parameters, and then compared these predictions quantitatively with empirical observations [8, 18]. Both of those studies used artificial manipulations of sensory reliability. We employed a natural cause of variation in visual estimates: the correlation between surface orientation and measurement error in estimating intersurface distance [1, 2]. Because this correlation is ubiquitous in everyday perception, observers in our study were more likely to use commonplace rather than ad hoc strategies. The fact that nearly optimal cue integration was observed in all three studies suggests that the phenomenon is pervasive.

The observed and predicted PSEs in our experiment were very similar (Figure 3D), but the observed and pre-

dicted JNDs differed consistently (Figures 4A and 4B). (A sign test for paired samples showed that the medians of observed and predicted JNDs in Figure 4C are different, $p = 0.007$, and that the medians of observed and predicted PSEs are not, $p = 0.435$; see Figure S1 in the Supplemental Data). If the PSEs and JNDs come from the same experimental measurements, how could one set of predictions match so closely while the other fell consistently short? There are at least two possibilities. First, the weights applied by the brain are themselves variable. And, second, the noises associated with the visual and haptic estimates are correlated.

Consider the first possibility. The ML rule assumes fixed weights for each experimental condition, but the weights in a biological system probably vary over time, even within a condition. If the weights were variable, but on average optimal (Equation 1), the PSEs would be unaffected because they are determined by the system's average response. Variable weights would, however, cause a JND increase because they are determined by trial-by-trial variability. In a Monte Carlo simulation, we determined how much weight variation would be needed to increase the JNDs to the observed values. We rewrite Equation 1:

$$\hat{S}_{VH} = N(w_V, \sigma_{wV}) \times N(\hat{S}_V, \sigma_V) + N(w_H, \sigma_{wH}) \times N(\hat{S}_H, \sigma_H), \quad (3)$$

where $N(\mu, \sigma)$ represents a normally distributed random variable with mean μ and standard deviation σ; w_V and w_H are the means, and σ_{wV} and σ_{wH} are the standard deviations of the weight distributions, respectively. We set σ_V and σ_H to the values measured in the within-modality experiment (JNDs in Figure 2) and set w_V and w_H according to Equation 1. We then found the values of σ_{wV} and σ_{wH} that increased JNDs to the observed values. With σ_{wH} and $\sigma_{wH} \approx 0.02$, the predicted and observed JNDs were equal, and the PSEs did not change.

Now consider the second possibility. When visual and haptic noises are correlated, the ML rule (Equation 1) yields less improvement in JNDs than predicted by Equation 2 [19], like we found. If the brain took the correlation into account and used appropriate weights [19], the PSEs would differ from the predictions of Equation 1, which would disagree with our results. However, if the brain did not take the correlation into account, and used the weights in Equation 1, the PSEs would be the same as we observed.

Thus, our data are consistent with the ML model (Equation 1), with small weight variation or with a small correlation that is not taken into account.

Experimental Procedures

Apparatus and Stimuli

The apparatus is described in [8]. Visual and haptic stimuli were two parallel planes with slants of 0°, 45°, or 90° (Figure 1; slant is defined, of course, relative to the line of sight). The head was stabilized with a chin-and-forehead rest. Stimulus distance from the eyes varied randomly (49–61 cm) to make distance to one surface an unreliable cue. Observers viewed the surfaces and/or grasped them with the index finger and thumb to estimate the intersurface distance. The visual and haptic stimuli were spatially aligned.

The visual stimuli were random-element stereograms. The simulated surfaces were textured with sparse rectangular elements (3-mm sides ± 1.5 mm, covering on average 5% of the surface) and were otherwise transparent. Because element size and density were randomized, they were not a reliable cue to intersurface distance. Surface areas were also randomized, so projected area and side overlap were also not useful cues. Textures were regenerated for each presentation. CrystalEyes shutter glasses were used to present different images to the two eyes. The refresh rate was 96 Hz.

The haptic stimuli were generated by using PHANToM force-feedback devices, one each for the index finger and thumb. The devices apply forces to the observer's digits to simulate the haptic experience of 3D objects. The digits were attached to the corresponding PHANToM with a thimble and elastic band. Observers were unaware of the thimbles and band during the experiment. 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 intersurface distance in the stimulus.

Observers

The same five observers with normal or corrected-to-normal vision participated in all experiments. Two were unaware of the experimental purpose.

Procedure

In the within-modality experiments, two stimuli, a standard and comparison, were presented in random order in each trial. Psychometric functions were measured with the method of constant stimuli (Figure 2). The intersurface distances were 50 mm for the standard and 44, 47, 49, 51, 53, or 56 mm for the comparison. Each pairing of standard and comparison was presented 30 times to each observer. Before each trial, the observer saw two 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 the cursors (representing the finger tips) disappeared. The disappearance was a signal to start pinching. In the haptics-alone condition, the observer felt two parallel (invisible) surfaces. The surfaces were extinguished 750 ms after both digits made contact. In the vision-alone condition, the pinch made both surfaces visible for 750 ms (no useful haptic cue was available). Trials consisted of two stimulus presentations. After the first one, the spheres reappeared, the observer inserted the digits, and the second presentation occurred. Observers indicated the stimulus with the apparently greater intersurface distance. No feedback was given.

In the intermodality experiment, two stimuli, conflict and no-conflict, were presented in random order on each trial. Psychometric functions were again measured with the method of constant stimuli (Figures 3A–3C). The visually and haptically specified distances in the no-conflict stimuli were equal and ranged from 44 to 56 mm. The visual and haptic distances in the conflict stimuli were {47, 53}, {48.5, 51.5}, {49.25, 50.75}, {50, 50}, {50.75, 49.25}, {51.5, 48.5}, or {53, 47} mm. Each pair was presented 30 times. Observers indicated which of the two stimuli contained the apparently greater intersurface distance. No feedback was given.

Supplemental Data

Supplemental Data including three figures are available at http://images.cellpress.com/supmat/supmatin.htm.

Acknowledgments

We thank Marc Ernst, James Hillis, Stanley Klein, Michael Kubovy, Michael Landy, Clifton Schor, and Dhanraj Vishwanath for valuable comments, Carmel Levitan for helping run the experiments, and Alison Dilworth for posing for Figure 1. Part of this work was presented at the Vision ScienceS Society meeting, 2002. This work was supported by grants from National Institutes of Health (EY12851), AFOSR (F49620-01-1-0417), and Silicon Graphics.

Received: October 7, 2002
Revised: November 11, 2002
Accepted: January 13, 2003
Published: March 18, 2003

References

1. McKee, S.P., Levi, D.M., and Bowne, S.F. (1990). The imprecision of stereopsis. Vision Res. *30*, 1763–1779.
2. van Beers, R.J., Wolpert, D.M., and Haggard, P. (2002). When feeling is more important than seeing in sensorimotor adaptation. Curr. Biol. *12*, 834–837.
3. Newport, R., Rabb, B., and Jackson, S.R. (2002). Noninformative vision improves haptic spatial perception. Curr. Biol. *12*, 1661–1664.
4. Clark, J.J., and Yuille, A.L. (1990). Data Fusion for Sensory Information Processing Systems (Boston: Kluwer).
5. Landy, M.S., Maloney, L.T., Johnston, E.B., and Young, M. (1995). Measuring and modeling of depth cue combination: in defense of weak fusion. Vision Res. *35*, 389–412.
6. Yuille, A.L., and Bülthoff, H.H. (1996). Bayesian decision theory and psychophysics. In Perception as Bayesian Inference, D.C. Knill and W. Richards, eds. (Cambridge: Cambridge University Press), pp. 123–161.
7. Ghahramani, Z., Wolpert, D.M., and Jordan, M.I. (1997). Computational models of sensorimotor integration. In Self-Organization, Computational Maps, and Motor Control, P.G. Morasso, and V. Sanguineti, eds. (Amsterdam: Elsevier), pp. 117–147.
8. Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.
9. Backus, B.T., Banks, M.S., van Ee, R., and Crowell, J.A. (1999). Horizontal and vertical disparity, eye position, and stereoscopic slant perception. Vision Res. *39*, 1143–1170.
10. Backus, B.T., and Banks, M.S. (1999). Estimator reliability and distance scaling in stereoscopic slant perception. Perception *28*, 217–242.
11. Banks, M., Hooge, T.C., and Backus, B. (2001). Perceiving slant about a horizontal axis from stereopsis. J. Vision *1*, 55–79.
12. Rogers, B.J., and Bradshaw, M.F. (1993). Vertical disparities, differential perspective and binocular stereopsis. Nature *361*, 253–255.
13. Rogers, B.J., and Bradshaw, M.F. (1995). Disparity scaling and the perception of frontoparallel surfaces. Perception *24*, 155–179.
14. Young, M.J., Landy, M.S., and Maloney, L.T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. Vision Res. *33*, 2685–2696.
15. Buckley, D., and Frisby, J.P. (1995). Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. Vision Res. *33*, 919–933.
16. Frisby, J.P., Buckley, D., and Horsman, J.M. (1995). Integration of stereo, texture, and outline cues during pinhole viewing of real ridge-shaped objects and stereograms of ridges. Perception *24*, 181–198.
17. van Beers, R.J., Sittig, A.C., and Denier van der Gon, J.J. (1999). Integration of proprioceptive and visual position-information: an experimentally supported model. Exp. J. Neurophysiol. *81*, 1355–1364.
18. Landy, M.S., and Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. J. Opt. Soc. Am. A *18*, 2307–2320.
19. Oruç, I., Maloney, L.T., and Landy, M.S. (2003). Weighted linear cue combination with possibly correlated error. Vision Res., in press.
20. Pugh, E.M., and Winslow, G.H. (1966). The analysis of physical measurements (Reading, MA: Addison-Wesley).