PERGAMON

Vision Research 41 (2001) 449–461

Vision Research

www.elsevier.com/locate/visres

Experience-dependent visual cue integration based on consistencies between visual and haptic percepts

Joseph E. Atkins, József Fiser, Robert A. Jacobs *

Department of Brain and Cognitive Sciences and the Center for Visual Science, University of Rochester, Rochester, NY 14627, USA

Received 5 May 2000; received in revised form 13 September 2000

Abstract

We study the hypothesis that observers can use haptic percepts as a standard against which the relative reliabilities of visual cues can be judged, and that these reliabilities determine how observers combine depth information provided by these cues. Using a novel visuo-haptic virtual reality environment, subjects viewed and grasped virtual objects. In Experiment 1, subjects were trained under motion relevant conditions, during which haptic and visual motion cues were consistent whereas haptic and visual texture cues were uncorrelated, and texture relevant conditions, during which haptic and texture cues were consistent whereas haptic and motion cues were uncorrelated. Subjects relied more on the motion cue after motion relevant training than after texture relevant training, and more on the texture cue after texture relevant training than after motion relevant training. Experiment 2 studied whether or not subjects could adapt their visual cue combination strategies in a context-dependent manner based on context-dependent consistencies between haptic and visual cues. Subjects successfully learned two cue combination strategies in parallel, and correctly applied each strategy in its appropriate context. Experiment 3, which was similar to Experiment 1 except that it used a more naturalistic experimental task, yielded the same pattern of results as Experiment 1 indicating that the findings do not depend on the precise nature of the experimental task. Overall, the results suggest that observers can involuntarily compare visual and haptic percepts in order to evaluate the relative reliabilities of visual cues, and that these reliabilities determine how cues are combined during three-dimensional visual perception. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Visual cue integration; Visual percepts; Haptic percepts; Relative reliability

1. Introduction

The visual environment provides many cues to visual depth, including cues based on binocular disparities, motion parallax, texture gradients, and shading. Experimental evidence indicates that human observers combine information provided by these cues when making depth judgments (e.g. Braunstein, 1968; Dosher, Sperling, & Wurst, 1986; Bruno & Cutting, 1988; Bülthoff & Mallot, 1988; Rogers & Collett, 1989; Nawrot & Blake, 1993; Landy, Maloney, Johnston, & Young, 1995). Moreover, this evidence suggests that observers' cue integration strategies are context-dependent; observers combine the information provided by the available cues in different ways depending on the current viewing conditions and goals of the observer.

It has been hypothesized that the extent to which an observer uses the information provided by a particular visual cue depends upon the estimated reliability of that cue relative to the estimated reliabilities of other cues (Maloney & Landy, 1989). This conjecture has received considerable empirical support. Johnston, Cumming, and Landy (1994) reported that subjects relied about equally on stereo and motion cues when making shape judgments at near viewing distances, whereas they relied more on the motion cue at far viewing distances. They argued that this context-dependency is sensible because stereo disparities are small at far viewing distances and, thus, small misestimates of disparity can lead to large errors in calculated depth. Related data was provided by Young, Landy, and Maloney (1993) who reported that when either a texture or motion cue was corrupted by added noise, subjects tended to rely more heavily on the uncontaminated cue when making depth judgments.

* Corresponding author. Tel.: +1-716-2750753; fax: +1-716-4429216.

E-mail address: robbie@bcs.rochester.edu (R.A. Jacobs).

0042-6989/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0042-6989(00)00254-6

If observers' cue integration strategies are based on the estimated relative reliabilities of the available visual cues, then this raises the issue of how observers are able to assess the relative reliabilities of these cues. For example, why do observers believe that motion and stereo cues are about equally reliable at signaling the depth of an object when the object is near to them, and on what basis do they conclude that stereo is a significantly less reliable cue to object depth when the object is far away?

At least part of the answer may be that observers compare the information provided by visual cues to the information provided by other sensory modalities. In particular, it has often been speculated that people learn how to visually perceive the world by comparing their visual percepts with percepts obtained during motor interactions with the environment. Historically, this idea may have been first proposed by Berkeley (1709/1910). Berkeley speculated that visual perception of depth results from associations between visual cues and sensations of touch and motor movement. More recently, Piaget (1952) used similar ideas to explain how children learn to interpret and attach meaning to retinal images based on their motor interactions with physical objects. Empirical data supporting the notion that motor interactions play a role in visual learning comes from prism adaptation studies in which subjects adapted to visual distortions produced by distorting lenses. Adaptation often occurs when subjects are allowed to interact with the environment (Held & Hein, 1958, 1963). In many studies subjects only became aware of the visual distortion through their motor interactions (Welch, 1978). For our own purposes, the most relevant experimental study is that of Ernst, Banks, and Bülthoff (2000) who found that subjects' estimates of visual slant relied more heavily on a visual cue when the cue was congruent with haptic feedback versus when it was incongruent with this feedback.

This article reports three experiments examining how observers develop their cue combination strategies for visual depth. In particular, we study the hypothesis that haptic percepts provide a standard against which the relative reliabilities of visual cues can be judged, and that these reliabilities determine how the cues are combined in order to achieve three-dimensional visual perception. The experiments used a novel visuo-haptic virtual reality environment which allowed observers not only to view virtual objects, but also to interact with them in a realistic manner. This environment was ideal for a cue-conflict experimental paradigm. The virtual reality apparatus allowed us to independently manipulate the depth indicated by each visual cue, and to independently manipulate the depth indicated by the haptic cue. Consequently, we were able to control the relative consistency between the haptic cue and each of the visual cues.

In all three experiments, subjects viewed and grasped vertically-oriented elliptical cylinders, and judged the depths of these cylinders. Visually, the cylinders were defined by motion and texture cues. In Experiment 1, subjects were trained under motion relevant conditions, meaning that motion and haptic cues were consistent (whereas texture and haptic cues were uncorrelated), and under texture relevant conditions, meaning that texture and haptic cues were consistent (and motion and haptic cues were uncorrelated). When subjects' visual cue combination strategies were examined, it was found that subjects relied more on the motion cue after motion relevant training than after texture relevant training, and more on the texture cue after texture relevant training than after motion relevant training. Experiment 2 studied whether or not subjects could adapt their visual cue combination strategies in a context-dependent manner on the basis of context-dependent consistencies between visual and haptic percepts. In one context, for example when the texture elements of a cylinder were red, the motion and haptic cues were consistent whereas the texture and haptic cues were inconsistent. This context is referred to as the motion relevant context. In a second context, for example when the texture elements were blue, the texture and haptic cues were consistent. This context is referred to as the texture relevant context. Trials belonging to motion relevant and texture relevant contexts were randomly intermixed. The results indicate that subjects successfully learned two cue combination strategies, and correctly applied each strategy in its appropriate context; they relied more on the motion cue in the motion relevant context than in the texture relevant context, and more on the texture cue in the texture relevant context than in the motion relevant context. In order to ensure that the results of the first and second experiments were not due to an idiosyncratic property of the experimental task, Experiment 3 replicated Experiment 1 except that it used a more naturalistic task. Because the same pattern of results was found in Experiment 1 and Experiment 3, we conclude that our findings are robust in the sense that they do not depend on the precise nature of the experimental task. Overall, we conclude that, consistent with the hypotheses of Berkeley, Piaget, and many others, observers can compare visual and haptic percepts in order to evaluate the relative reliabilities of visual cues. Moreover, these reliabilities determine how the cues are combined during three-dimensional visual perception.

2. General methods

2.1. Experimental apparatus

The visuo-haptic virtual reality experimental apparatus consisted of virtual reality goggles and two PHAN-

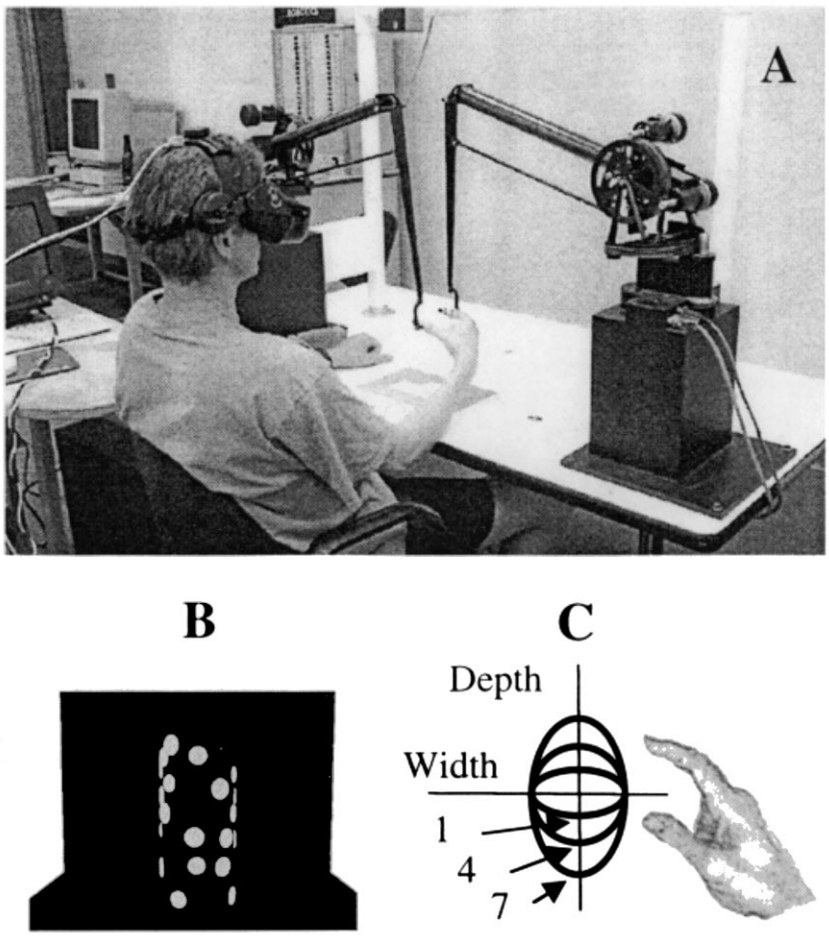

Fig. 1. (A) A subject using the visuo-haptic virtual reality experimental apparatus. The subject is grasping a virtual object viewed via displays embedded in the head-mounted goggles. (B) A typical instance of the display that the subjects viewed during the experiment. The motion cue cannot be illustrated, but the texture cue is evident from the foreshortening of the disks at the sides of the cylinder. (C) A schematic representation of the cylinders viewed from the top. The three ellipses represent three of the possible seven cylinder shapes (1 = smallest depth; 4 = depth equal to width; 7 = largest depth).

ToM™ 3D Touch interfaces that were attached by two fingerholders to the subject's thumb and index fingers (see Fig. 1, Panel A). This apparatus allowed subjects to physically interact with virtual objects viewed via the goggles in a natural way using a wide range of movements (e.g. grasping, moving, or throwing objects). The 3D Touch interfaces generated force fields that created haptic sensations (e.g. weight, hardness, and friction) appropriate to the motor interactions with the object displayed in the goggles. The apparatus also allowed for independent manipulation of the visual and haptic cues regarding these objects.[1]

[1] Technical details regarding the experimental apparatus are available on the world wide web at www.sensable.com/products/phantom.htm.

2.2. Stimuli

The stimuli were vertically-oriented elliptical cylinders (cylinders whose horizontal cross-sections are ellipses). The horizontal cross-section of a cylinder may have been circular, in which case the cylinder was equally deep as wide, may have been elliptical with a principal axis parallel to the observers' line of sight, in which case the cylinder was more deep than wide, or may have been elliptical with a principal axis parallel to the frontoparallel plane, in which case the cylinder was less deep than wide. The height of a cylinder was 150 mm; the width of a cylinder was 60.5 mm. The depth of a cylinder took one of seven possible values; these values were evenly spaced in the range between 35.75 and 85.25 mm (see Fig. 1, Panel C).

Haptically, the cylinders were defined by haptic sensations obtained when subjects grasped the cylinders using their thumb and index fingers. Subjects' hands were not visible during a grasp. Three markings at the

top of the visual display helped subjects orient their finger positions. One marking was fixed; it indicated the location of the center of a cylinder. The other two markings showed the position of the two fingers along the width axis. Subjects were instructed to grasp the cylinder so that the three markings overlapped; this occurred when the fingers were oriented along the depth axis. Although subjects found it easy to orient their fingers in the requested manner, conditions were established so that the haptic cue to a cylinder's depth was invariant to the orientation of a subject's fingers.

Visually, the cylinders were defined by texture and motion cues. Subjects viewed the cylinders monocularly from an orthogonal perspective (the cylinders' sides were visible but not their tops or bottoms; see Fig. 1, Panel B). Conditions were established so as to eliminate the possibility that subjects could obtain information about the depth of a cylinder based on head movements. The viewing angle was fixed so that the horizontal component of an observer's line of sight was parallel to the depth axis regardless of the observer's head movements (this prevented subjects from looking 'behind' the cylinder). In addition, the distance from the observer to the center of the cylinder was fixed at 406 mm.

The texture and motion cues were created through the use of flat 'disks' that were placed along the surface of a cylinder, and that traveled horizontally along this surface. The number of disks was proportional to the surface area of a cylinder; the initial position of each disk and the size of each disk was randomized with the constraint that there was minimal overlap among disks. The two-dimensional image of the disks contained gradients of texture element density, size, and compression which were texture cues to the shape of a cylinder (see Fig. 1, Panel B). Previous studies have shown that gradients of texture element compression are the primary (nearly exclusive) determinants of observers' perceptions of depth or shape for the types of stimuli used here (Cutting & Millard, 1984; Blake, Bülthoff, & Sheinberg, 1993; Cumming, Johnston, & Parker, 1993; Knill, 1998). The motion cue was created by the relative horizontal motions of the disks along the cylinder surface. The velocity of the motion was constant within a display; it was randomized between displays. Note that the cylinder did not rotate; rather the disks moved along the surface of static cylinders. Thus, the stimuli were different from kinetic depth effect stimuli which were not used because they produce artifactual depth cues when the horizontal cross-section of a cylinder is non-circular, such as changes in retinal angle subtended by the cylinder over time. The motion cue in the stimuli used here is an instance of a constant flow field. Constant flow fields produce reliable and robust perceptions of depth (Perotti, Todd, & Norman, 1996; Perotti, Todd, Lappin, & Phillips, 1998).

The experiments used a cue-conflict experimental paradigm in which the cylinder depths indicated by haptic, texture, and motion cues were independently manipulated. The computer graphics manipulation used to create the cue conflict between texture and motion cues was nearly identical to the one presented by Young et al. (1993), and is described in detail in Jacobs and Fine (1999). In short, for each visual display two cylinders of identical heights and widths, but different depths, were defined. One cylinder was used to create the texture cue, and the other cylinder was used to create the motion cue. The cylinders were positioned so that their midpoints lay at the origin of a three-dimensional coordinate system. Parallel projection was used to map the coordinates of a location on one cylinder to the coordinates of the corresponding location on the other cylinder. Consequently, it was possible for a texture element to have its compression at each point in time determined by the shape of one cylinder, but its motion at each point in time determined by the shape of the other cylinder. Observers perceived only one object, even though the texture elements conveyed two object shapes: one shape was indicated by the texture element compressions, and the other shape by the texture element motions.

2.3. Procedure

Experiments consisted of training trials and test trials. On each training trial in Experiments 1 and 2, subjects had unlimited time to visually and haptically inspect the depth of a cylinder that was located at the center of the workspace. After inspecting the cylinder, subjects moved their hands to the workspace periphery, and were then forced to relate the visual and haptic cues to a cylinder's depth by requiring them to perform a cross-modal same/different judgment task. If the subject believed that visual and haptic percepts indicated cylinders of the same depth, then they responded 'same'; otherwise they responded 'different'. Subjects then received a visual signal indicating whether their response was correct or incorrect. A large cube appeared which covered the workspace center; if the response was correct, the color of the cube was green; if the response was incorrect, the color was red. Importantly, the subjects were asked to judge the consistency between the haptic cue and the overall visual perception of depth rather than the depth indicated by any individual visual cue. In addition, subjects were not aware that the environment contained independent motion and texture cues.

Unbeknownst to the subjects, training trials could be classified as either motion relevant or texture relevant. As a matter of notation, define set M to be the collection of displays in which the cylinder shape indicated by the motion cue was one of the seven possible shapes, and in which the shape indicated by the texture cue was circular (the cylinder was equally deep as wide). Define set T to be the collection of displays in which texture indicated one of the seven possible shapes, whereas motion indicated a circular shape. On motion relevant training trials,

the visual display was a member of set M. On trials in which the subject was informed that the visual and haptic cues indicated cylinders of the same depth, the cylinder shape indicated by the haptic cue was identical to the shape indicated by the motion cue, whereas the shapes indicated by haptic and texture cues were uncorrelated. Thus only the motion cue provided information that was useful for performing the experimental task under motion relevant training conditions. Similarly, during texture relevant training trials, the visual display was a member of set T. On trials in which the subject was informed that the visual and haptic cues were consistent, the cylinder shape indicated by the haptic cue was identical to the shape indicated by the texture cue, and the shapes indicated by haptic and motion cues were uncorrelated. In this case, only the texture cue provided information that was useful for performing the experimental task.

It is important to understand the nature of the experimental task. The feedback provided to subjects regarding the correctness of their same/different judgments did not directly inform them as to how to adapt their visual cue combination strategies. This information could only be obtained by relating visual and haptic percepts. In addition, the experimental task was designed so as to encourage subjects to adapt their visual cue integration strategies, and to discourage them from adapting their interpretations of individual visual cues, a form of learning known as cue recalibration. The information provided to subjects was not conducive to the adaptation of either depth-from-motion estimates or depth-from-texture estimates. Consider, for example, motion relevant training trials in which the subject was informed that the haptic and visual cues were consistent. In this case, haptic and motion cues signaled the same depth, meaning that the motion cue was already properly calibrated. The texture cue, on the other hand, should not be recalibrated because it was uncorrelated with the haptic cue (and with the motion cue), meaning that there was no information suggesting that depth-from-texture estimates ought to be either smaller or larger. Analogous remarks apply to texture relevant training trials. Although the possibility that subjects showed some degree of cue recalibration cannot strictly be ruled out, we believe that the experimental results described below are best interpreted as consistent with the hypothesis that subjects showed experience-dependent adaptation of their visual cue integration strategies.[2]

Two types of test trials were used in the experiments, motor test trials and visual test trials. Subjects did not receive feedback on test trials. The test trials were designed to permit an estimation of subjects' cue combination strategies. In particular, we wanted to estimate the relative degree to which a subject relied on the motion cue versus the texture cue when making visual depth judgments about displays that contained both cues. For this purpose, it was assumed that observers linearly combine depth information based on motion and texture cues:

$$d(m, t) = w_M d(m) + w_t d(t) \quad (1)$$

where m and t denote the motion and texture cues respectively, $d(m, t)$ is the percept of visual depth based on both cues, $d(m)$ is the depth percept based on the motion cue, dt is the depth percept based on the texture cue, and w_M and w_T are the linear coefficients corresponding to the motion and texture cues (it was also assumed that w_M and w_T are non-negative and sum to one). Linear cue combination rules are often assumed in the visual perception literature, and they have received a considerable degree of empirical support (e.g. Dosher et al., 1986; Bruno & Cutting, 1988; Landy et al., 1995). We found that a linear combination rule provides a good fit to the experimental data reported in this article. To complete the specification of Eq. (1), it is necessary to specify observers' depth perceptions based on the motion cue, $d(m)$, and based on the texture cue, $d(t)$. Because there is no uncontroversial method for estimating these values, and for the sake of simplicity, we assumed that the depth estimates based on these cues are each veridical. The veridical assumption is approximately correct, and is commonly made by researchers studying cue combination rules (e.g. Tittle, Norman, Perotti, & Phillips, 1997; van Ee, Banks, & Backus, 1999).

On *motor* test trials, subjects performed a cross-modal matching task during which they viewed a display of a cylinder and positioned their thumb and index fingers so as to indicate the cylinder's perceived depth. Motor test trials either used displays from set M or displays from set T.[3] At the start of a trial, a large, blue cube covered the entire workspace center. This cube then disappeared, revealing a cylinder. A subject had unlimited time to view the cylinder, then reached into the center of the workspace and held his thumb and index fingers at the perceived cylinder depth for 1000

[2] The issue of whether changes in responses to multiple-cue stimuli are due to changes in observers' cue combination strategies or to changes in observers' interpretations of individual cues has been problematic for many studies. For the sake of simplicity, other investigators have typically referred to the underlying cause as changes in observers' cue combination strategies (e.g. Ernst, Banks, Bülthoff, 2000; van Ee, Banks, Backus, 1999).

[3] In Experiments 1 and 3, a block of motor test trials following motion relevant training used cylinder displays from set M, and used displays from set T following texture relevant training. In Experiment 2, half of the motor test trials in a block were presented in a motion relevant context and used displays from set M, and half the trials were presented in a texture relevant context and used displays from set T.

ms during which time their response was measured. No parts of the subject's body were visible in the display, and no haptic percepts were provided to the subject. After making a response, the cube appeared again and the subject moved his hand to the workspace periphery. Based on the linear cue combination rule, it was possible to apply linear regression to each subject's responses on the motor test trials in order to obtain maximum likelihood estimates, using a Gaussian likelihood function, of that subject's motion and texture weights. The regression function had only one free parameter, namely the motion coefficient w_M (recall that $w_T = 1 - w_M$).

On *visual* test trials, subjects performed a two-alternative forced-choice task during which they viewed two successively displayed cylinders and judged which cylinder was greater in depth. Because the display of one cylinder was from set M whereas the display of the other cylinder was from set T, visual test trials allowed us to assess the relative degree to which a subject relied on the motion cue versus the texture cue when making visual depth judgments. At the start of a trial, a large, blue cube covered the workspace center. This cube then disappeared, revealing a cylinder for 2000 ms. Next, the cube reappeared for 1000 ms, followed by a second cylinder for 2000 ms. The subject then judged which cylinder was greater in depth. Subjects did not grasp cylinders or receive haptic percepts during visual test trials. For the purpose of estimating a subject's cue weights, it was assumed that the subject used the linear cue combination strategy to obtain depth estimates for the cylinders depicted in each display, and then used a probabilistic rule in order to select the display depicting the deeper cylinder. We assumed that the probabilistic rule could be approximated using a logistic function (a monotonic, differentiable function whose shape resembles a multidimensional 'S'). In short, the rule considers the difference between the perceived depths of the cylinders depicted in displays M and T, and then uses a logistic function to map this difference to a probability. If the difference is positive, then the observer is more likely to choose display M as depicting the deeper cylinder; if the difference is negative, then the observer is more likely to choose display T; if the difference is zero, then the observer is equally likely to choose either display (mathematical details of this probabilistic model are given in Jacobs & Fine, 1999). Based on the linear cue combination strategy and the probabilistic rule, we applied logistic regression to each subject's responses on the visual test trials in order to obtain maximum likelihood estimates, using a Bernoulli likelihood function, of that subject's motion and texture weights. The regression function had two free parameters, namely the motion coefficient w_M and a temperature parameter τ which determines the overall steepness of the logistic surface.

2.4. Subjects

Subjects were students at the University of Rochester. They had normal or corrected-to-normal vision. They were naive to the purposes of the experiments.

3. Experiment 1

Experiment 1 studied differences in observers' visual cue combination rules after prolonged experience under the motion relevant condition (haptic and motion cues were correlated) versus after prolonged experience under the texture relevant condition (haptic and texture cues were correlated). Four of the seven subjects initially performed training trials under the motion relevant condition followed by motor and visual test trials, and then performed training trials under the texture relevant condition followed by motor and visual test trials. The order of conditions was counterbalanced across subjects (the remaining subjects were trained and tested in the reverse order: first texture relevant training and testing, then motion relevant training and testing). Our prediction was that subjects would adapt their visual cue combination strategies so that they relied more on the motion cue after motion relevant training than after texture relevant training, and more on the texture cue after texture relevant training than after motion relevant training.

Subjects performed two blocks of training trials (under motion relevant training conditions for example) on the first three days of participation in the experiment, where a block consisted of 84 trials. On Day 3, they also performed a block of motor test trials (42 trials) and a block of visual test trials (98 trials). On Days 4–5, subjects performed a block of training trials, two blocks of visual test trials, and two blocks of motor test trials. Days 6–10 were identical to Days 1–5 except that the relevant visual cue on the training trials was reversed (texture relevant training, for example).[4]

The results for one subject, subject JH, on the visual test trials are shown in Fig. 2. Recall that each visual test trial included a display from set M and a display from set T. Consequently, four values are needed to represent the stimulus conditions on any trial: the depths indicated by the motion and texture cues in the display from set M, and the depths indicated by these cues in the display from set T. However, because the texture cue in the display from set M and the motion

[4] The description of the schedule of training and test trials for Experiments 1–3 is accurate for a typical subject. In some cases, deviations from this schedule occurred either because a subject showed especially slow learning performance, and thus was provided with extra training trials, or because of equipment failure.

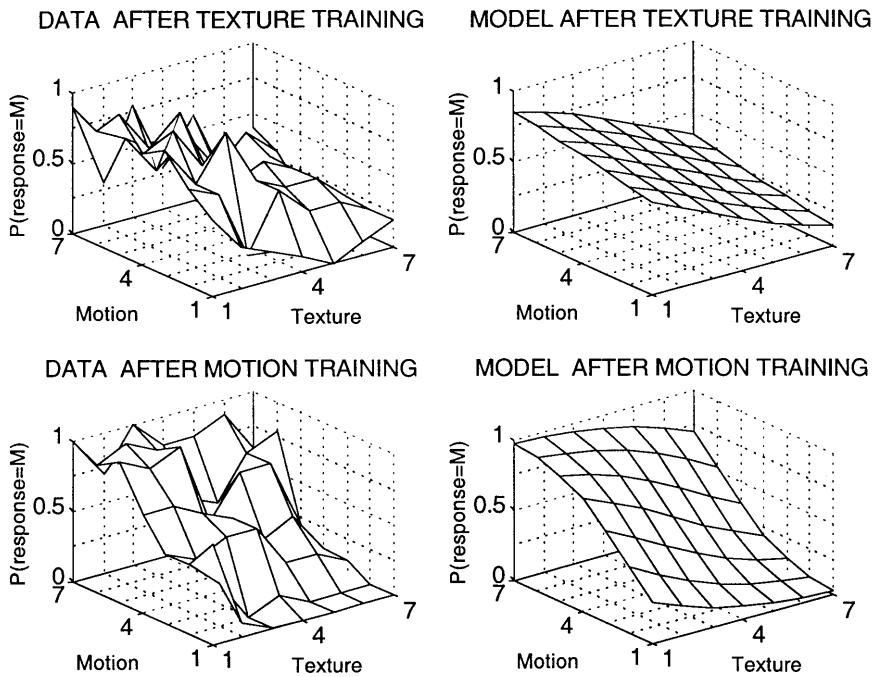

Fig. 2. The response data of subject JH on visual test trials following texture relevant training (top-left graph) and motion relevant training (bottom-left graph). The logistic model was used to fit surfaces to these two datasets (top-right and bottom-right graphs, respectively).

cue in the display from set T always indicated a circular cylinder, these constant values can be omitted and, thus, the stimulus conditions can be represented by two values. The axis labeled 'Motion' in each graph in Fig. 2 gives the depth indicated by the motion cue in the display from set M ($1 =$ smallest depth; $7 =$ greatest depth). The axis labeled 'Texture' gives the depth indicated by the texture cue in the display from set T. The axis labeled 'P (response $= M$)' gives the probability that the subject chose the display from set M as depicting the deeper cylinder.

Subject JH was initially trained under the texture relevant condition; this training was followed by motion relevant training. The top-left graph of Fig. 2 gives this subject's response data on the visual test trials following texture relevant training. The shape of this graph is intuitively sensible. As the motion cue in the display from set M indicated a deeper cylinder (that is, as the value along the motion axis increases), it became more likely that the subject picked display M as depicting a deeper cylinder. Similarly, as the texture cue in the display from set T indicated a deeper cylinder (as the value along the texture axis increases), it became less likely that the subject picked display M as depicting a deeper cylinder. The top-right graph shows a logistic surface that was fit to the subject's response data based upon the probabilistic model described above. Analogous graphs for the test trials following motion relevant training are shown in the bottom of Fig. 2. The bottom-left graph shows the subject's response data; the bottom-right graph shows the logistic surface that was fit to this data.

A comparison of the graphs in the top and bottom rows of Fig. 2 reveals that the subject responded to the same set of test trials in different ways following texture relevant and motion relevant training conditions. The gradient of the response data (or of the logistic surface) along the Texture axis is greater following texture relevant training than it is following motion relevant training. This means that the subject relied more on the texture cue following texture relevant training than following motion relevant training. Similarly, the gradient of the response data along the motion axis is greater following motion relevant training than it is following texture relevant training, meaning that the subject relied more on the motion cue following motion relevant training than following texture relevant training. On the basis of this data, we conclude that this subject adapted her visual cue combination strategy in an experience-dependent manner based on the consistencies (and inconsistencies) between haptic and visual cues.

Fig. 3 shows the results of visual and motor tests for all seven subjects who participated in Experiment 1. The horizontal axis identifies a subject; the vertical axis gives the estimated value of a subject's motion coefficient w_M. The light bars and the dark bars indicate the motion coefficient based on the test trials following motion relevant training and following texture relevant training, respectively. Based on the visual test trials, all seven subjects had larger motion weights following

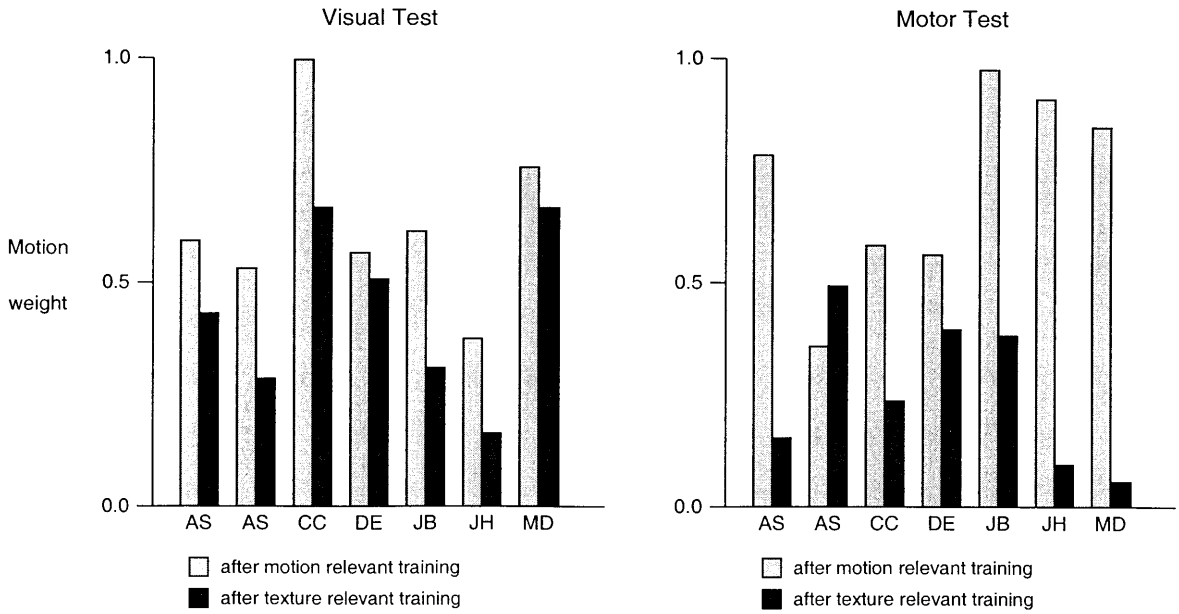

Fig. 3. The estimated motion coefficient for each subject following motion relevant and texture relevant training based on visual and motor test trials.

motion relevant training than following texture relevant training (see the graph on the left). Define the motion coefficient difference to be the estimated value of w_M after motion relevant training minus its estimated value after texture relevant training. The average motion coefficient difference is 0.2 (the standard error of the mean is 0.039) which is significantly greater than zero ($t = 5.15$, $P < 0.002$ based on a one-tailed t-test). The results based on motor test trials are very similar (see the graph on the right). With a single exception, all subjects had larger motion weights after motion relevant training than after texture relevant training. The average motion coefficient difference is 0.46 (standard error = 0.133), which is significantly greater than zero ($t = 3.46$, $P < 0.013$).

In conclusion, the results of Experiment 1 support the experimental hypothesis that haptic percepts provide a standard against which the relative reliabilities of visual cues can be judged, and that these reliabilities determine how the cues are combined. When motion and haptic cues are consistent and texture and haptic cues are uncorrelated, observers seem to (unconsciously) conclude that motion is a more reliable cue than texture. Consequently, they adjust their visual cue combination rules so to emphasize the depth information provided by motion and to discount the information provided by texture. Under the opposite conditions, when texture and haptic cues are consistent but motion and haptic cues are uncorrelated, observers conclude that the texture cue is more reliable and adjust their cue combination rules so as to emphasize texture-based information and to discount motion-based information.

4. Experiment 2

In order to accurately estimate depth under various visual conditions, our visual systems need to use different cue combination strategies in different contexts. Experiment 2 evaluated whether or not observers can use context-dependent consistencies between visual and haptic percepts in order to learn and apply two different context-dependent visual cue combination strategies. If haptic and motion cues are consistent in one context, and haptic and texture cues are consistent in another context, will observers adapt their cue combination rules so as to emphasize depth-from-motion estimates in the first context and depth-from-texture estimates in the second context?

The experiment was identical to Experiment 1 with the following exceptions. Whereas Experiment 1 had separate stages for motion relevant and texture relevant training, Experiment 2 contained only a single stage. Unbeknownst to the subjects, half of the trials in Experiment 2 belonged to a motion relevant context and the remaining trials belonged to a texture relevant context. During a training trial belonging to the motion relevant context, the visual display was a member of set M, and the texture elements were rendered in a specific color, such as red. When a subject was informed that visual and haptic percepts indicated cylinders of the same depth, the cylinder shape indicated by the haptic cue was identical to the shape indicated by the motion cue, but uncorrelated with the shape indicated by the texture cue. Consequently, only the motion cue provided useful information for performing the cross-

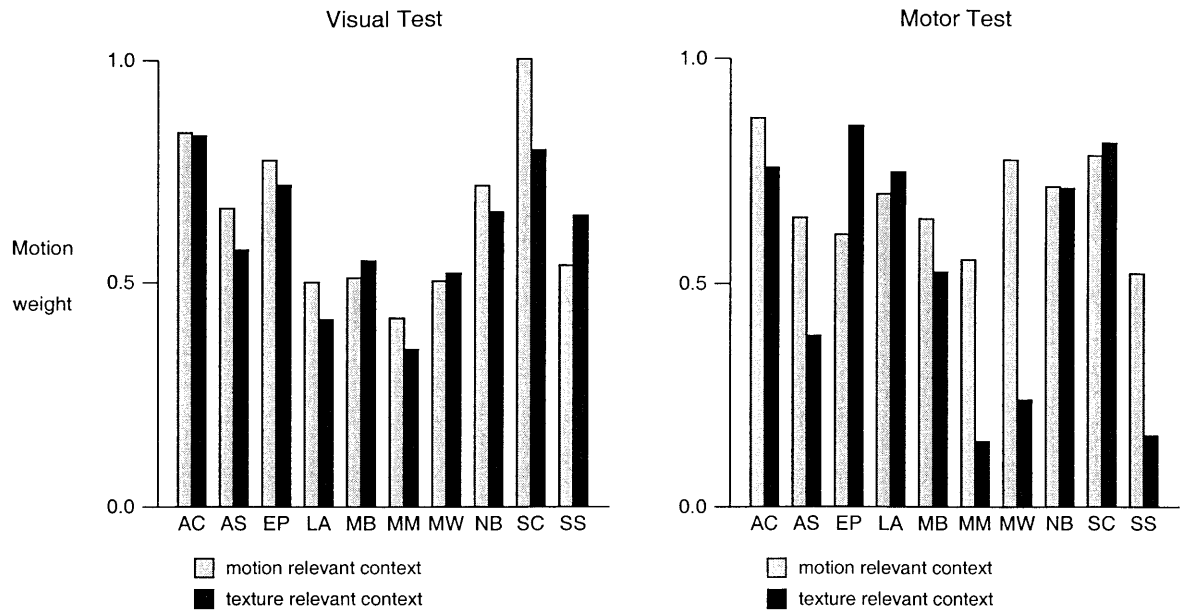

Fig. 4. The estimated motion coefficient for each subject in the motion relevant and texture relevant contexts based on visual and motor test trials.

modal same/different judgment task. In order to do well on this task, the subject needed to learn that when he or she is viewing a cylinder with red texture elements, then depth-from-motion information should be emphasized. In contrast, during a texture relevant training trial, the visual display was a member of set T, and the texture elements were rendered in another color, such as blue. When a subject was informed that visual and haptic percepts indicated cylinders of the same depth, the cylinder shapes indicated by texture and haptic cues were identical, whereas the shapes indicated by motion and haptic cues were uncorrelated. In this case, the subject needed to learn that when he or she is viewing a cylinder with blue texture elements, then depth-from- texture information should be emphasized. The relationship between color (red versus blue) and context (motion relevant versus texture relevant) was counterbalanced across subjects.

Subjects participated in the experiment for 8 days. On Days 1–6, they performed two blocks of training trials, where a block consisted of 84 trials. On Day 6, they also performed a block of motor test trials (56 trials) and a block of visual test trials (98 trials). On Days 7–8, subjects performed a block of training trials, two blocks of motor test trials, and two blocks of visual test trials. Training blocks were organized into 4 groups of 21 trials; groups alternated between trials belonging to the motion relevant context and trials belonging to the texture relevant context. Importantly, however, during test blocks, trials belonging to the motion relevant or texture relevant context were randomly intermixed.

The results of Experiment 2 are shown in Fig. 4. Ten subjects participated in the experiment. Their estimated motion weights in the motion relevant context (light bars) and in the texture relevant context (dark bars) based on the visual test trials are shown in the graph on the left; the graph on the right gives their motion weights in each context based on the motor test trials. We first discuss the results of the visual test trials. Seven of the ten subjects had larger motion weights in the motion relevant context than in the texture relevant context. Define the motion coefficient difference to be the difference in the value of a subject's motion weight in the motion relevant context versus the texture relevant context. The average motion coefficient difference is 0.04 (standard error $= 0.027$) which is marginally significantly greater than zero ($t = 1.496$, $P = 0.084$). In regard to the data based on the motor test trials, seven of the ten subjects had larger motion weights in the motion relevant context. The average motion coefficient difference is 0.148 (standard error $= 0.076$) which is significantly greater than zero ($t = 1.94$, $P < 0.05$). On the basis of this data, we conclude that subjects adapted their visual cue combination strategies so as to emphasize depth-from-motion information in the context in which motion and haptic cues were consistent, and to emphasize depth-from-texture information in the context in which texture and haptic cues were consistent.

As discussed in the introduction, previous investigators have shown that observers' visual cue combination strategies are flexible in the sense that they are context-dependent; i.e. these strategies make greater or lesser use of different cues in different visual contexts. For example, Johnston et al. (1994) reported that subjects relied about equally on stereo and motion cues when making shape judgments at near viewing distances,

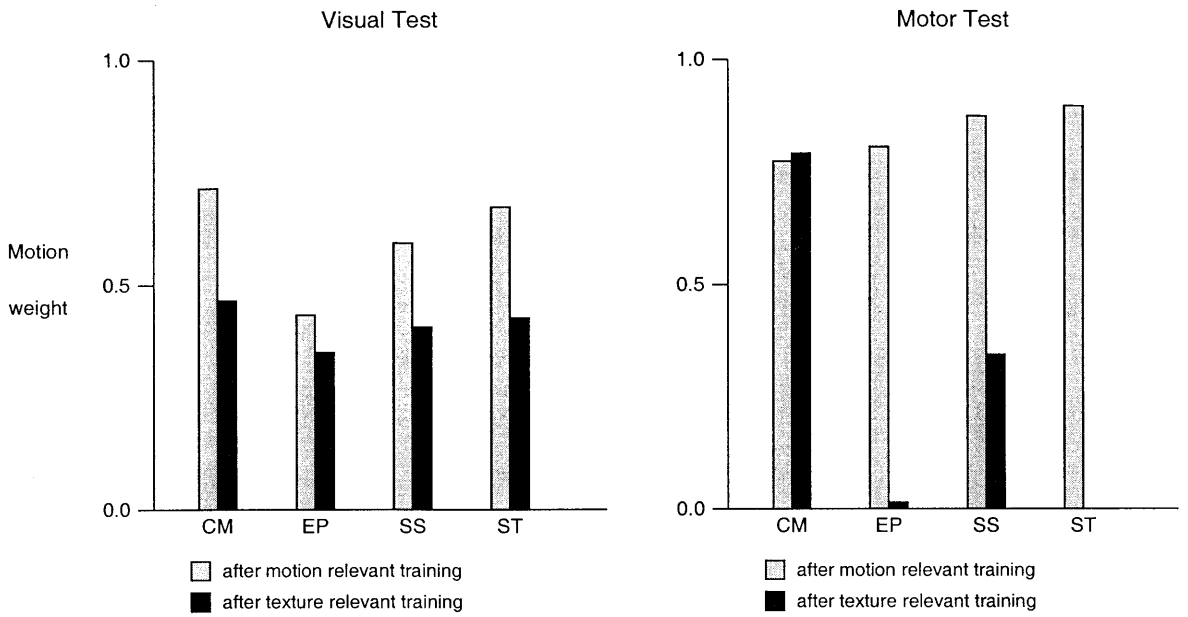

Fig. 5. The estimated motion coefficient for each subject following motion relevant and texture relevant training based on visual and motor test trials.

whereas they relied more on the motion cue at far viewing distances. The results of Experiment 2 suggest that observers can use context-dependent consistencies between visual and haptic percepts in order to learn context-dependent visual cue combination strategies.

5. Experiment 3

Training trials in Experiments 1 and 2 used a cross-modal same/different judgment task with feedback. Because it could be argued that the use of feedback is not 'naturalistic', Experiment 3 replicated Experiment 1 except that its training trials used a different procedure. This procedure did not include feedback; instead it relied on the fact that observers both viewed and grasped cylinders. This procedure was close to a typical everyday situation in which a person obtains visual and haptic percepts of the depth of an object, such as a drinking cup, when the person views and then grasps the object.

During a training trial in Experiment 3, subjects first performed a cross-modal matching task during which they viewed a display of a cylinder and positioned their thumb and index fingers so as to indicate the cylinder's perceived depth. Next, they grasped the cylinder along the depth axis, thereby obtaining a haptic cue to the cylinder's depth. Finally, subjects judged whether their cross-modal estimate of depth based on the visual cues was greater than, less than, or the same as the depth indicated by the haptic cue. Subjects were asked to make this judgment in order to force them to relate visual and haptic percepts. Importantly, subjects did not receive feedback about the correctness of their judgments. As before, training trials could be classified as motion relevant or texture relevant. During a motion relevant trial, the visual display was a member of set M, and motion and haptic cues indicated cylinders of the same depth (depths indicated by texture and haptic cues were uncorrelated). During a texture relevant trial, the display was a member of set T, and texture and haptic cues were consistent. Half of the subjects were first trained under motion relevant conditions followed by texture relevant conditions. This order was reversed for the remaining subjects.

On the first day in which subjects participated in the experiment, subjects performed two blocks of training trials (under motion relevant conditions, for example), where a block consisted of 42 trials. On Days 2–4, subjects completed three blocks. Subjects performed two blocks of training trials, one block of motor test trials (28 trials), and one block of visual test trials (98 trials) on Day 5, and one block of training trials, two blocks of motor test trials, and two blocks of visual test trials on Day 6. Days 7–12 were identical to Days 1–6 except that the relevant visual cue on the training trials was reversed (texture relevant training, for example).

Fig. 5 shows the results of visual (left graph) and motor (right graph) tests for all four subjects who participated in the experiment. The light and dark bars give the estimated motion coefficient based on test trials following motion relevant and following texture relevant training, respectively. Based on the visual test trials, all four subjects had larger motion weights fol-

lowing motion relevant training than following texture relevant training. Define the motion coefficient difference to be the estimated value of the motion weight after motion relevant training minus its estimated value following texture relevant training. The average motion coefficient difference is 0.193 (standard error = 0.039) which is significantly greater than zero ($t = 4.963$, $P < 0.01$). In regard to the motor test trials, three of the four subjects had larger motion weights following motion relevant training. The average motion coefficient difference is 0.551 (standard error = 0.205) which is significantly greater than zero ($t = 2.69$, $P < 0.05$).

Similar to the results of Experiment 1, the results of Experiment 3 support the hypothesis that haptic percepts provide a standard against which the relative reliabilities of visual cues can be evaluated. Moreover, these reliabilities determine how the cues are combined. Taken in conjunction with the results of Experiment 1, these results also suggest that our findings are robust in the sense that they do not depend on the precise nature of the experimental task.[5]

6. Summary and conclusions

This article has addressed the issue of how observers are able to estimate the relative reliabilities of the available cues in a visual environment. Good estimates are important because these estimates are used by observers in order to integrate information provided by different cues into a unified percept. Berkeley (1709/1910), Piaget (1952), and many others, speculated that people learn to visually perceive the world by comparing their visual percepts with percepts obtained during motor interactions with the environment. We have studied the hypothesis that haptic percepts can provide a standard against which the relative reliabilities of different visual cues can be estimated, and that these relative reliabilities determine how the cues are combined in order to achieve three-dimensional visual perception. In Experiment 1, it was found that subjects relied more on a motion cue after motion relevant training than after texture relevant training, and more on a texture cue after texture relevant training than after motion relevant training. Experiment 2 studied whether or not subjects could adapt their visual cue combination strategies in a context-dependent manner based on context-dependent consistencies between haptic and visual cues. The results indicate that subjects successfully learned two cue combination strategies simultaneously, and correctly applied each strategy in its appropriate context. Experiment 3 was similar to Experiment 1 except that it used a more naturalistic experimental task in the sense that the only signals provided to subjects were haptic and visual percepts. Because the same pattern of results was found in Experiments 1 and 3, the findings do not depend on the precise nature of the experimental task. Overall, the results of these experiments suggest that observers can involuntarily compare visual and haptic percepts in order to evaluate the relative reliabilities of visual cues, and that these reliabilities determine how the cues are combined.

Although the idea that people learn to visually perceive the world by comparing their visual percepts with percepts obtained during motor interactions has existed for a long time, this hypothesis has been difficult to study. It is arguably the case that the experiments reported here and the recent work of Ernst et al. (2000) are the most direct and detailed empirical evaluations of this hypothesis. Using visual displays that contained stereo and texture cues to slant, Ernst et al. found that subjects' estimates of visual slant relied more heavily on a visual cue when that cue was congruent with haptic feedback, a result that is in qualitative agreement with our own results. These two studies suggest that the use of haptic percepts to estimate the reliabilities of visual cues is general in the sense that it can be demonstrated under a variety of experimental conditions, and with respect to a variety of visual cues and visual judgments. Our experiments also show that observers can use context-dependent consistencies between haptic and visual percepts in order to learn multiple cue combination strategies. We believe that this finding will play an important role in future theories that attempt to explain the complexity, flexibility, and robustness of observers' visual depth judgments in natural settings.

The reported experiments raise a number of issues that will need to be examined in future studies. For example, we need to know the neural site and mechanism for the adaptation of observers' visual cue integration strategies. Previous investigators hypothesized that the primate visual system is organized into two independent pathways, referred to as either the 'what' and 'where' pathways (Ungerleider & Mishkon, 1982) or the 'what' and 'how' pathways (Milner & Goodale, 1995). The 'what' pathway is a ventral stream that computes visual object properties (such as object shape and depth), whereas the 'where' or 'how' pathway is a dorsal stream that computes spatial properties necessary for sensorimotor control (such as positional properties needed to grasp an object). Because haptic percepts obtained during grasping influenced observers' visual depth judgments, we speculate that the adapta-

[5] In all the experiments reported here it is typically the case that subjects' data on the visual and motor tests are very similar. However, there are exceptions to this rule. In Experiment 3, for instance, subjects CM and ST show similar results on the visual test but dissimilar results on the motor test. Understanding the relationships between the responses required by visual and motor tests and understanding the nature of individual differences in subjects' responses are important challenges for future studies.

tion found in our experiments occurs at an early stage of visual processing that precedes the separation into ventral and dorsal pathways and that produces outputs which are used by both pathways, or else that it occurs in the dorsal pathway but that these changes are able to influence visual judgments typically associated with the ventral pathway. Additional support for these possibilities is the fact that qualitatively similar results were found based on visual and motor test trials, suggesting that a common (or perhaps tightly coupled) set of computations underlie visual depth judgments regardless of whether or not a task requires a motor response.

In regard to a neural mechanism underlying the adaptation, we speculate that this mechanism can be characterized as a distributed gain-control process similar to those found in other modulatory mechanisms such as the distance-dependent mechanisms found along the monkey dorsal and ventral pathways (e.g. Sakata, Shibutani, & Kawano, 1980; Colby, Duhamel, & Goldberg, 1993; Gnadt & Mays, 1995; Dobbins, Jeo, Fiser, & Allman, 1998; Trotter & Celebrini, 1999). This type of neural computation is a good candidate because it has been reported to operate at several stages of the visual system, and because it can implement a variety of important neural properties such as the invariant visual responses of cells found in the ventral pathway, and the coordinate transformations thought to be computed via neural gain fields of cells in the dorsal pathway (Salinas & Abbott, 1996, 1997).

Additionally, we need to know more about the relationships between visual perception and motor interactions. The experimental results reported here suggest that subjects regarded haptic percepts as providing 'ground truth' information about object depth. There are at least three possible reasons why this was the case. First, haptic cues may have a privileged status relative to other cues such that subjects are biased towards believing that haptic percepts are veridical. This possibility is unlikely to be correct, however, because previous investigators have demonstrated circumstances in which shape judgments are closest to the shape indicated by visual cues when visual and haptic cues are in conflict (Rock & Victor, 1964). Second, subjects may have regarded haptic percepts as veridical based on correlational information; the haptic cue was positively correlated with one of the visual cues, whereas the other visual cue was uncorrelated with all other cues. Third, subjects may have unconsciously noticed that the displays contained visual cue conflicts and, thus, concluded that haptic percepts were reliable whereas visual percepts were questionable. In general, haptic and visual cues can be compared when objects are nearby but not when they are far away. When objects are far away, observers may use motor interactions other than grasping in order to learn about the reliabilities of visual cues. Future studies should assess whether or not signals based on self-motion, accommodation, or vergence are useful for evaluating visual cue reliabilities.

In conclusion, we have provided evidence that observers' visual cue integration strategies are dynamically modified in response to changing cue reliabilities as signaled by haptic percepts. We also showed that observers can learn more than one cue combination strategy simultaneously, and that they can apply each strategy in its appropriate context. These results suggest a plausible framework for how observers learn to compute visual depth from multiple cues in an accurate, flexible, and robust manner. These findings also support theories of infant development which suggest that motor interactions play an important role in the acquisition of aspects of visual perception (Bushnell & Boudreau, 1993; Bertenthal, 1996).

Acknowledgements

We thank R. Aslin, M. Banks, and D. Knill for helpful discussions of this material, and M. von der Heyde for creating the experimental environment. We also thank R. Aslin and D. Williams for commenting on an earlier version of this manuscript, D. Knill for important suggestions regarding Experiment 3, and D. Ballard and M. Hayhoe for allowing us to use their lab facilities. This work was supported by NIH grants R29-MH54770, R01-EY13149, and P41-RR09283, by NSF Graduate Fellowship DGE-9616170, and by McDonnell-Pew Program in Cognitive Neuroscience grant JSMF-96-32.

References

Berkeley, G. (1709/1910) Essay towards a new theory of vision. London: Dutton.

Bertenthal, B. I. (1996). Origins and early development of perception, action, and representation. *Annual Review of Psychology*, *47*, 431–459.

Blake, A., Bülthoff, H. H., & Sheinberg, D. (1993). Shape from texture: ideal observers and human psychophysics. *Vision Research*, *33*, 1723–1737.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology*, *117*, 161–170.

Braunstein, M. L. (1968). Motion and texture as sources of slant information. *Journal of Experimental Psychology*, *78*, 247–253.

Bushnell, E. W., & Boudreau, J. P. (1993). Motor development and the mind: the potential role of motor abilities as a determinant of aspects of perceptual development. *Child Development*, *64*, 1005–1021.

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1993). Ventral intraparietal area of the macaque: anatomic location and visual response properties. *Journal of Neurophysiology*, *69*, 902–914.

Cumming, B. G., Johnston, E. B., & Parker, A. J. (1993). Effects of different texture cues on curved surfaces viewed stereoscopically. *Vision Research*, *33*, 827–838.

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America*, *5*, 1749–1758.

Cutting, J. E., & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General*, *113*, 198–216.

Dobbins, A. C., Jeo, R. M., Fiser, J., & Allman, J. M. (1998). Distance modulation of neural activity in the visual cortex. *Science*, *281*, 552–555.

Dosher, B. A., Sperling, G., & Wurst, S. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. *Vision Research*, *26*, 973–990.

Ernst, M. O., Banks, M. S., & Bülthoff, H. H. (2000). Touch can change visual slant perception. *Nature Neuroscience*, *3*, 69–73.

Gnadt, J. W., & Mays, L. E. (1995). Neurons in monkey parietal area LIP are tuned for eye-movement parameters in 3-dimensional space. *Journal of Neurophysiology*, *73*, 280–297.

Held, R., & Hein, A. (1958). Adaptation to disarranged hand-eye coordination contingent upon reafferent stimulation. *Perceptual and Motor Skills*, *8*, 87–90.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*, 872–876.

Jacobs, R. A., & Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. *Vision Research*, *39*, 4062–4075.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of motion and stereopsis cues. *Vision Research*, *34*, 2259–2275.

Knill, D. C. (1998). Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture. *Vision Research*, *38*, 2635–2656.

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

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. *Visual communications and image processing IV: proceedings of the SPIE*, *1199*, 1154–1163.

Milner, D. A., & Goodale, M. A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.

Nawrot, M., & Blake, R. (1993). On the perceptual identity of dynamic stereopsis and kinetic depth. *Vision Research*, *33*, 1561–1571.

Perotti, V. J., Todd, J. T., Lappin, J. S., & Phillips, F. (1998). The perception of surface curvature from optical motion. *Perception and Psychophysics*, *60*, 377–388.

Perotti, V. J., Todd, J. T., & Norman, J. F. (1996). The visual perception of rigid motion from constant flow fields. *Perception and Psychophysics*, *58*, 666–679.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Rock, I., & Victor, J. (1964). Vision and touch: an experimentally created conflict between the two senses. *Science*, *143*, 594–596.

Rogers, B. J., & Collett, T. S. (1989). The appearance of surfaces specified by motion parallax and binocular disparity. *The Quarterly Journal of Experimental Psychology*, *41*, 697–717.

Sakata, H., Shibutani, H., & Kawano, K. (1980). Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey. *Journal of Neurophysiology*, *43*, 1654–1672.

Salinas, E., & Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Academy of Sciences USA*, *93*, 11956–11961.

Salinas, E., & Abbott, L. F. (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology*, *77*, 3267–3272.

Tittle, J. S., Norman, J. F., Perotti, V. J., & Phillips, F. (1997). The perception of scale-dependent and scale-independent surface structure from binocular disparity, texture, and shading. *Perception*, *26*, 147–166.

Trotter, Y., & Celebrini, S. (1999). Gaze direction controls response gain in primary visual cortex neurons. *Nature*, *398*, 239–242.

Ungerleider L.G., & Mishkin M. (1982). Two cortical visual systems. In: D.J. Engle, M.A. Goodale, & R.J. Mansfield (Eds.), *Analysis of Visual Behaviour*, (pp. 549–586), Cambridge, MA: MIT Press.

van Ee, R., Banks, M. S., & Backus, B. T. (1999). An analysis of binocular slant contrast. *Perception*, *28*, 1121–1145.

Welch, R. B. (1978). *Perceptual modification: adapting to altered sensory environments*. New York: Academic Press.

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, *33*, 2685–2696.

Further reading

Heller M.A. (1992). Haptic dominance in form perception: Vision versus proprioception, *Perception*, *21*, 655–660.