PERGAMON

Vision Research 39 (1999) 3621–3629

Vision Research

www.elsevier.com/locate/visres

Section 4

Optimal integration of texture and motion cues to depth

Robert A. Jacobs *

Center for Visual Science, University of Rochester, Rochester, NY 14627, USA

Received 2 December 1998; received in revised form 31 March 1999

Abstract

We report the results of a depth-matching experiment in which subjects were asked to adjust the height of an ellipse until it matched the depth of a simulated cylinder defined by texture and motion cues. In one-third of the trials the shape of the cylinder was primarily given by motion information, in another one-third of the trials it was given by texture information, and on the remaining trials it was given by both sources of information. Two optimal cue combination models are described where optimality is defined in terms of Bayesian statistics. The parameter values of the models are set based on subjects' responses on trials when either the motion cue or the texture cue was informative. These models provide predictions of subjects' responses on trials when both cues were informative. The results indicate that one of the optimal models provides a good fit to the subjects' data, and the second model provides an exceptional fit. Because the predictions of the optimal models closely match the experimental data, we conclude that observers' cue-combination strategies are indeed optimal, at least under the conditions studied here.

Keywords: Depth; Optimal integration; Texture; Motion

1. Introduction

The human visual system obtains information about depth from a large number of cues. Cues to depth result from object rotation, observer motion, binocular vision in which the two eyes receive different patterns of light, texture gradients in retinal images, and many other factors (Cutting & Vishton, 1995). Recently, there has been a significant increase in the number of studies examining strategies observers use to combine information provided by each of multiple cues in a visual environment (e.g. Dosher, Sperling & Wurst, 1986; Bruno & Cutting, 1988; Bülthoff & Mallot, 1988; Rogers & Collett, 1989; Johnston, Cumming & Parker, 1993; Nawrot & Blake, 1993; Young, Landy & Maloney, 1993; Landy, Maloney, Johnston & Young, 1995; Tittle, Norman, Perotti & Phillips, 1997; Turner, Braunstein & Andersen, 1997; Jacobs & Fine, 1999).

This article addresses the question of whether or not observers' cue combination rules for visual depth can be characterized as optimal in a Bayesian statistical sense described below. This question is important for several reasons. If observers' judgments are optimal in a particular context, then this suggests that their perceptual systems are operating in a principled manner within that context, and that we can use our definition of optimality in order to reasonably conjecture as to what those principles are. For instance, an optimal model based on Bayes' rule may make assumptions about the visual environment. If this model provides a good fit to observers' judgments, then it is reasonable to hypothesize that observers may also be making those same assumptions. Furthermore, if observers' behaviors result from a learning process, and those behaviors are optimal, then this places a strong constraint on hypotheses regarding the underlying learning mechanisms. Proposed learning mechanisms that result in optimal behaviors are viable hypotheses, whereas mechanisms that do not result in such behaviors are not.

Some researchers have conjectured that observers' cue combination rules may indeed be optimal (e.g. Landy et al., 1995), but there is relatively little available empirical data that directly evaluates this hypothesis in a quantitative manner. A notable exception is the recent work of Knill (1998). He asked subjects to make

* Tel.: +1-716-275-0753; fax: +1-716-442-9216.
E-mail address: robbie@bcs.rochester.edu (R.A. Jacobs)

0042-6989/99/$ - see front matter
PII: S0042-6989(99)00088-7

judgments about planar surface orientations based on multiple texture cues. By comparing subjects' responses to predictions of different optimal models (referred to as 'ideal observers'), he was able to estimate the weight given to each cue by subjects, and also the strengths of various assumptions that subjects made about texture information.

Most investigations assess observers' combination rules in a way that does not address the issue of whether or not these rules are optimal. Sometimes these investigations make assumptions that preclude the possibility of evaluating optimality. For instance, Tittle et al. (1997) had subjects make shape judgments about stimuli containing binocular disparity, texture, and shading cues. In order to study subjects' cue combination rules, they assumed that subjects' shape estimates based on individual cues were veridical, and that these veridical estimates were linearly combined. Although these assumptions made it easy to estimate the linear coefficients that subjects used when combining information, they also made it impossible to assess whether or not the subjects' cue combination rules were optimal. Other investigations have shown that observers' combination rules are sensible, but not that they are necessarily optimal. For example, Young et al. (1993) used a perturbation technique in order to analyze depth perceptions based on texture and motion cues. They found that when either cue was corrupted by added noise, subjects tended to rely more heavily on the uncontaminated cue. While this result suggests that observers combination rules are sensible, the experiment does not provide sufficient detail in order to assess whether or not these rules are statistically optimal.

The present article reports the results of a depth-matching experiment in which subjects were asked to adjust the height of an ellipse until it matched the depth of a simulated cylinder defined by texture and motion cues. In one-third of the trials the shape of the cylinder was primarily given by motion information, in another one-third of the trials it was given by texture information, and in the remaining trials it was given by both sources of information. Two optimal cue combination models are described where optimality is defined in terms of Bayesian statistics. The parameter values of the models are set based on subjects' responses on trials when either the motion cue or the texture cue was informative. These models provide predictions of subjects' responses on trials when both cues were informative. The results indicate that one of the optimal models provides a good fit to the subjects' data, and the second model provides an exceptional fit. The results are surprising because the models are strongly constrained (they are linear), and because the first model has no free parameters whereas the second model has only one free parameter. Because the predictions of the optimal models closely match the experimental data, we conclude that observers' cue combination strategies are indeed optimal, at least under the conditions studied here. Section 2 describes the two optimal models. Section 3 describes the experiment. In Section 4, the statistical metrics of bias and variance are used to analyze the experimental results. The bias of a subject's response indicates whether the subject tended to overestimate or underestimate the depth of a cylinder. The variance of a subject's response measures the amount of variability in the response. The subjects' data are also compared to the predictions of the optimal cue combination models.

2. Optimal cue combination models

We define the optimal estimate of visual depth given motion and texture cues as the depth, denoted d, that maximizes the probability $P(d|m, t)$ where m and t denote the motion and texture cues. Using Bayes' rule, this probability may be re-written as

$$P(d|m, t) \propto P(m, t|d)P(d) \tag{1}$$

Assuming that the motion and texture cues are conditionally independent given the depth, we arrive at the equation

$$P(d|m, t) \propto P(m|d)P(t|d)P(d) \tag{2}$$

where, using Bayes' rule,

$$P(m|d) = \frac{P(d|m)P(m)}{P(d)} \tag{3}$$

$$P(t|d) = \frac{P(d|t)P(t)}{P(d)} \tag{4}$$

The first optimal cue combination model that we consider assumes that the prior probability distributions of the depth, $P(d)$, of the motion cue, $P(m)$, and of the texture cue, $P(t)$, are uniform (meaning that all possible depths, all possible motion cues, and all possible texture cues are equally likely). Consequently, we refer to this model as Optimal Model-Uniform, henceforth referred to as model OM-U. In this case,

$$P(d|m, t) \propto P(d|m)P(d|t) \tag{5}$$

Note that the probability of depth d factors into the product of two terms: the first term is the probability of d given just the motion cue, and the second is the probability of d given just the texture cue. We assume that the probability distributions $P(d|m)$ and $P(d|t)$ are Normal distributions. Let d^*_m denote the optimal estimate of depth given just the motion cue [this is the depth that maximizes $P(d|m)$], and let d^*_t denote the optimal estimate of depth given just the texture cue [$d^*_t = \text{argmax}_d \; P(d|t)$]. If $d^*_m \approx d^*_t$, then Yuille and Bülthoff (1996) showed that the optimal estimate of depth based on both cues, denoted d^*, is given by

$$d^* = w_m d^*_m + w_t d^*_t \quad (6)$$

where

$$w_m = \frac{\sigma_m^{-2}}{\sigma_m^{-2} + \sigma_t^{-2}} \quad (7)$$

$$= \frac{\sigma_t^2}{\sigma_t^2 + \sigma_m^2} \quad (8)$$

and

$$w_t = \frac{\sigma_t^{-2}}{\sigma_m^{-2} + \sigma_t^{-2}} \quad (9)$$

$$= \frac{\sigma_m^2}{\sigma_t^2 + \sigma_m^2} \quad (10)$$

and σ_m^2 and σ_t^2 are the variances of the distributions $P(d|m)$ and $P(d|t)$ respectively. This solution has several appealing properties. First, the optimal estimate of depth based on both motion and texture cues is a linear combination of the optimal estimates based on the individual cues. Second, the linear coefficients, the weights w_m and w_t, are non-negative and sum to one. Third, the weight on a cue, such as the motion weight w_m, is large when that cue is relatively reliable (the variance σ_m^2 is smaller than the variance σ_t^2), and small when the cue is relatively unreliable (σ_m^2 is larger than σ_t^2). Eq. (6) is among the simplest optimal cue combination rules. For reasons that will become clear below, it is useful to also define a slightly more complicated optimal model. In this new model, it is assumed that observers are biased toward believing that objects are approximately equally deep as wide (e.g. the horizontal cross-section of objects is circular) when considering individually either a motion cue or a texture cue, though this assumption is not used when considering both cues. Under the experimental conditions described below, this 'circularity' assumption is consistent with the compactness assumption proposed by Caudek and Proffitt (1993), which is an instance of what those authors referred to as a perceptual heuristic. We refer to the optimal cue combination rule using the circularity assumption as Optimal Model-Circular, henceforth referred to as model OM-C. Computationally, the circularity assumption is implemented by making the prior probability distribution $P(d)$ in Eqs. (3) and (4) be a Normal distribution with mean d^*_P and variance σ_P^2, where the value of the mean is set equal to the width of the object (the distribution $P(d)$ in Eq. (2) is, as before, a uniform distribution). In this case,

$$P(d|m, t) \propto \frac{P(d|m)P(d|t)}{2P(d)} \quad (11)$$

If $d^*_m \approx d^*_t \approx d^*_P$, then it can be shown that the optimal estimate of depth based on both cues is given by (cf. Yuille & Bülthoff, 1996)

$$d^* = w_m d^*_m + w_t d^*_t - 2w_P d^*_P \quad (12)$$

where

$$w_m = \frac{\sigma_m^{-2}}{\sigma_m^{-2} + \sigma_t^{-2} - 2\sigma_P^{-2}} \quad (13)$$

$$w_t = \frac{\sigma_t^{-2}}{\sigma_m^{-2} + \sigma_t^{-2} - 2\sigma_P^{-2}} \quad (14)$$

$$w_P = \frac{\sigma_P^{-2}}{\sigma_m^{-2} + \sigma_t^{-2} - 2\sigma_P^{-2}} \quad (15)$$

The optimal estimate of depth based on both motion and texture cues is a linear combination of the optimal estimates based on the individual cues and the optimal estimate based on the prior distribution of depths.

Eqs. (6) and (12) are two optimal cue combination rules. An experiment was conducted to evaluate how well these rules predict observers' responses on a depth-matching task.

3. General methods

3.1. Stimuli and apparatus

The stimuli consisted of elliptical cylinders whose shapes were simulated on a 2-D video display by appropriate texture and motion algorithms. The horizontal cross-section of a cylinder could be circular, in which case the cylinder is equally deep as wide, could be elliptical with a principal axis parallel to the observer's line of sight (and minor axis parallel to the frontoparallel plane), in which case the cylinder is more deep than wide, or could be elliptical with a principal axis parallel to the frontoparallel plane (and minor axis parallel to the observers' line of sight), in which case the cylinder is less deep than wide. Twenty cylinder shapes were used in the experiment. The height (320 pixels) and width (160 pixels) of the cylinders were constant; only the simulated depths of the cylinder shapes varied. The 20 shapes had simulated depths that were equally spaced in the interval ranging from 80 to 270 pixels.

Three types of stimuli were defined: texture-informative stimuli; motion-informative stimuli; and texture-and-motion informative stimuli. In the texture-informative stimuli, the texture cue was created by mapping a homogeneous and isotropic texture consisting of circular spots to the surface of each cylinder using a texture mapping algorithm (the details of this algorithm are described in Hearn & Baker, 1997). Circular spots were placed on a two-dimensional sheet whose width was equal to the circumference of a horizontal cross-section of the cylinder, and whose height was equal to the height of the cylinder. The radius of each spot was randomly sampled from a uniform distribution ranging from 10 to 16 pixels. Either 65 or 80 spots were placed on the sheet, and the placement of the spots was random with the restriction that spots

could not overlap. The texture mapping algorithm mapped the sheet to the surface of the cylinder (though not to the top and bottom of the cylinder which were never visible to the observer). When a three-dimensional curved surface is projected onto a two-dimensional image, changes in surface orientation result in gradients of texture element size, shape, and density in the image. These gradients are texture cues to the shape of a cylinder.

The motion-informative stimuli were created as follows. Small points of light were initially placed on the surface of a simulated cylinder. Each point of light was a circle whose radius was 2 pixels. Either 65 or 80 points were placed on the cylinder, and the initial placement of the points was random with the restriction that points were about as far apart as the centers of the texture elements in the texture-informative stimuli. A movie was created by moving the points horizontally along the simulated surface of a cylinder in either a clockwise or anticlockwise direction. Speaking metaphorically, the motion of a point may be regarded as analogous to the motion of a train traveling around a track; the shape of the track is given by the circumference of a cylinder's horizontal cross-section. The velocity of the points was constant within a stimulus presentation; this velocity was varied between presentations. Points traveled the circumference of a cylinder's horizontal cross-section in either 55 or 75 frames. Note that the cylinder did not rotate; rather, the points moved along the simulated surface of static cylinders. Thus, the stimuli were different from kinetic depth effect (KDE) stimuli (except when the horizontal cross-section of a cylinder was circular, in which case the stimuli were identical to KDE stimuli). KDE stimuli were not used because they produce artifactitious 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 (e.g. Perotti, Todd, Lappin & Phillips, 1998; Perotti, Todd & Norman, 1996).[1]

[1] It is perhaps worth noting that whereas the texture-informative stimuli contained only texture cues to the shape of a cylinder (gradients of texture element size, shape, and density), the motion-informative stimuli did not contain only motion cues. Rather, the gradient of the density of the points of light in any frame of a motion-informative stimulus is a type of texture cue. Why, then, do we believe that the motion-informative stimuli are appropriately named? Because the density gradient is an extremely weak cue to a cylinder's shape, it is reasonable to believe that, at least to a first approximation, information in the motion-informative stimuli regarding this shape is largely or exclusively carried by the motions of the points of light. Other researchers have also made this assumption (e.g. Perotti et al., 1998). Evidence that density gradients are a weak cue to shape comes from multiple sources (e.g. Blake, Bülthoff & Sheinberg, 1993; Cumming, Johnston & Parker, 1993; Cutting & Millard, 1984; Knill, 1998).

The texture-and-motion informative stimuli contained both texture and motion cues to the shape of a cylinder. The appearances of the texture elements in each frame of a movie were rendered using the texture mapping algorithm described above; the motions of the texture elements were simulated using the motion algorithm described above.

The visual image of a cylinder subtended 2.21° of visual angle in the horizontal dimension and 4.6° in the vertical dimension. Stimuli were viewed monocularly from a distance of 1.45 m. They were rendered using a PowerComputing 225 computer (a clone of an Apple Macintosh) and a Sony Trinitron Multiscan 20sf II monitor. The video format was 100 Hz, noninterlace. The background luminance was 0.02 cd/m^2 and the luminance of the texture elements or the points of light was 30 cd/m^2.

3.2. Procedure

The experimental task was a depth-matching task in which subjects were asked to adjust the height of an ellipse until it matched the depth of a simulated cylinder (see Fig. 1). On each trial, subjects viewed a cylinder as depicted in either a texture-informative, motion-informative, or texture-and-motion informative stimulus. The center of the stimulus appeared 260 pixels to the left of the center of the video screen. The stimulus was displayed for 1000 ms, then it was erased, and then it was displayed again 333 ms later. This pattern was repeated until the subject made a response.

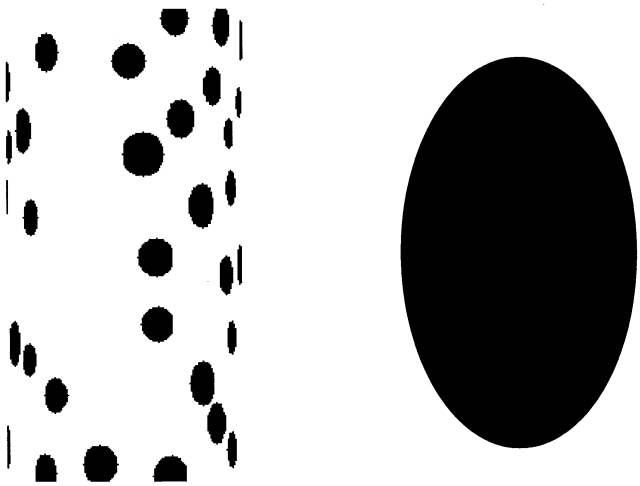

Fig. 1. On each trial, an experimental stimulus appeared on the left side of the video screen, and an ellipse appeared in the center of the screen. The subject could increase or decrease the height of the ellipse by pressing the 'd' and 'k' keys on the keyboard. The subject was instructed to adjust the height of the ellipse so that it matched the depth of the cylinder depicted in the experimental stimulus. The left side of this figure shows an instance of a texture-informative stimulus. The height of the ellipse on the right side of the figure is equal to the depth of the simulated cylinder depicted in the stimulus.

In addition to the experimental stimulus, an ellipse also appeared at the start of a trial. The width of the ellipse was equal to the width of the cylinders (160 pixels); the height of the ellipse was initially 170 pixels. The ellipse appeared at the center of the video screen. During the trial, the subject could increase or decrease the height of the ellipse by pressing the 'd' and 'k' keys on the keyboard. The subject was instructed to adjust the height of the ellipse so that it matched the depth of the cylinder depicted in the experimental stimulus. When the subject believed that the ellipse's height matched the cylinder's depth, he or she pressed the return key. (Note that if the height of the ellipse exactly matched the depth of the cylinder, then the shape of the ellipse was identical to the shape of a horizontal cross-section of the cylinder.) The experimental stimulus was then erased. On training trials, the subject received feedback; the 'target' ellipse was displayed 260 pixels to the right of the center of the video screen for 2333 ms. The height of the target ellipse was equal to the depth of the depicted cylinder. The word 'Response' appeared below the ellipse that the subject adjusted, and the word 'Target' appeared below the target ellipse. On test trials, the subject was not shown the target ellipse.

Subjects participated in the experiment for 6 days. The 6 days generally occurred within a 2-week period. On Days 1–3, subjects completed six blocks of training trials (this took about 1 h). Each block contained 60 trials (each of the 20 cylinders was depicted in each of the texture-informative, motion-informative, and texture-and-motion informative stimulus conditions in a random order). On Days 4–6, subjects completed two blocks of training trials followed by five blocks of test trials.

3.3. Subjects

The three subjects were graduate students at the University of Rochester. They had normal or corrected-to-normal vision. They were naive to the purposes of the experiment.

4. Results

In the statistics literature, it is common to characterize a statistical estimator using the metrics of bias and variance. We use these same metrics to characterize the subjects' responses on the depth-matching task. The bias of a subject's response at depth d is defined as

$$\text{bias}(d) = \langle \hat{d} \rangle - d \tag{16}$$

where $\hat{d}$ is a subject's depth estimate when shown a stimulus depicting a cylinder whose true depth is d, and the brackets $\langle\ \rangle$ denote an average. The bias is positive if the subject tended to overestimate the true depth of a cylinder, and negative if the subject tended to underestimate this depth. The variance of a subject's response at depth d is defined as

$$\text{variance}(d) = \langle (\hat{d} - \langle \hat{d} \rangle)^2 \rangle \tag{17}$$

Using the definition of the mean squared error (MSE) of a subject's response at depth d as

$$\text{MSE}(d) = \langle (\hat{d} - d)^2 \rangle \tag{18}$$

it is easy to show that the mean squared error can be expressed as the sum of two terms, one involving the bias and the other involving the variance, as follows (e.g. Casella & Berger, 1990)

$$\text{MSE}(d) = (\text{bias}(d))^2 + \text{variance}(d) \tag{19}$$

Characterizing the subjects' responses on the experimental task using the metrics of bias and variance is sensible for our current purposes. Recall that the optimal cue combination rules defined above are dependent on the variances of subjects' responses. That is, the variances of a subject's responses when using texture-informative stimuli and when using motion-informative stimuli help determine the optimal responses when using texture-and-motion informative stimuli as discussed above (see Eqs. (6)–(10) and (12)–(15)). In addition, according to the second optimal cue combination model described above (model OM-C) the biases of a subject's responses also help determine the optimal responses because they indicate the nature of a subject's prior distribution over cylinder depths (see Eqs. (12)–(15)). These points are illustrated below.

The 3×3 array of graphs in Fig. 2 shows the results of the experiment for three subjects. Each column corresponds to a different subject. The horizontal axis of each graph gives the depth of a cylinder in pixels. The vertical axes of the graphs in the top row give the bias of a subject's response on the test trials for each depth; the standard deviation of a subject's response is given in the graphs in the middle row; the root mean squared error (RMSE) of a subject's response is given in the graphs in the bottom row. The dotted line in each graph is for responses to motion-informative stimuli; the dashed line is for responses to texture-informative stimuli; the solid line is for responses to texture-and-motion informative stimuli. Note that, using the relationship in Eq. (19), the square of the data in the bottom row (the square of the RMSEs is the MSEs) is equal to the square of the data in the top row (the square of the biases) plus the square of the data in the middle row (the square of the standard deviations is the variances).

The data in Fig. 2 reveal many features of subjects' responses. The biases in subjects' responses tended to be positive when subjects were viewing cylinders that are less deep than wide, and negative when they were viewing cylinders that are more deep than wide. This

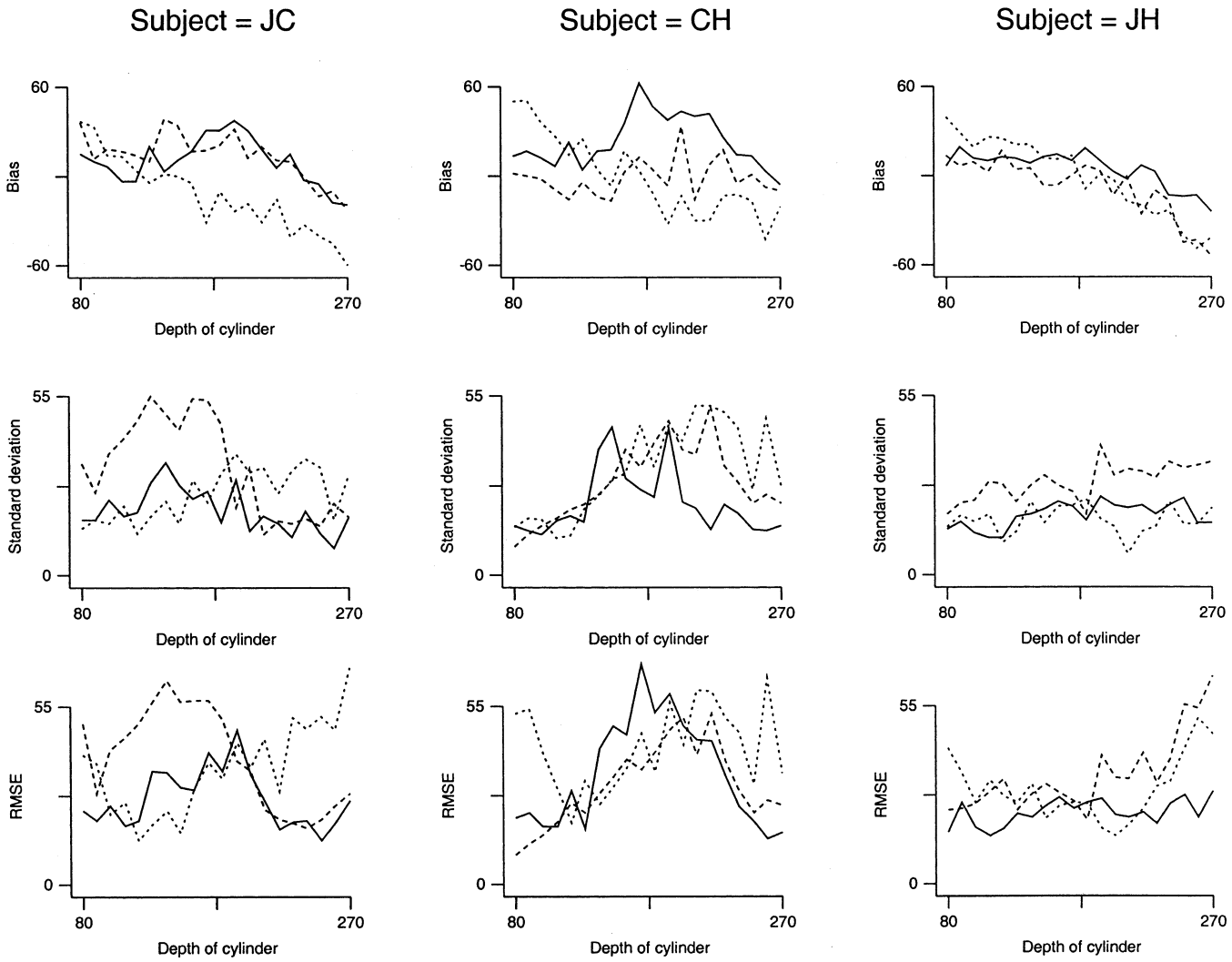

Fig. 2. The results of the experiment. Each column corresponds to a different subject. The graphs in the top row give the bias of a subject's response for each cylinder depth; the standard deviation of a subject's response is given in the graphs in the middle row; the root mean squared error (RMSE) of a subject's response is given in the graphs in the bottom row. The dotted line in each graph is for responses to motion-informative stimuli; the dashed line is for responses to texture-informative stimuli; the solid line is for responses to texture-and-motion informative stimuli.

trend is most evident for subjects JC and JH, and appears to be strongest when these subjects were viewing texture-informative and motion-informative stimuli, and less strong when these subjects were viewing texture-and-motion informative stimuli. This observation suggests that when subjects were viewing stimuli that contained only one informative cue, they tended to assume that the cylinders were roughly circular. However, when viewing stimuli that contained both cues, they either did not make this assumption, or else they made it less strongly. It is possible that subjects may have adopted this 'circularity' assumption because the average depth of the cylinders used in the experiment was nearly equal to the cylinders' width. Overall, subjects' responses were less variable when both cues were available. The standard deviations of the subjects' responses tended to be less when they were viewing stimuli that contained both texture and motion cues compared to when they were viewing stimuli that contained only a texture cue or only a motion cue. Consistent with the above observations regarding biases and variances, it was also the case that subjects' depth judgments were more accurate when both texture and motion cues were present in a stimulus. Errors tended to be larger under single-cue viewing conditions than under multiple-cue viewing conditions.

We compared the subjects' responses on test trials using the texture-and-motion stimuli to those predicted by the optimal cue combination models defined above. The parameter values of the optimal models were set based on subjects' responses on trials when only one cue was informative. Let optimal(d) denote an optimal response; this is the predicted average response of a subject to a texture-and-motion stimulus depicting a

cylinder of depth d. Let $\langle \hat{d}_m \rangle$ and $\langle \hat{d}_t \rangle$ denote the subject's average responses to motion-informative and texture-informative stimuli depicting a cylinder of depth d, respectively. Using the first optimal cue combination model (model OM-U), the optimal response is equal to

$$\text{optimal}(d) = w_m \langle \hat{d}_m \rangle + w_t \langle \hat{d}_t \rangle \quad (20)$$

where the linear coefficients w_m and w_t were computed based on the variances of the subject's responses to motion-informative and texture-informative stimuli according to Eqs. (7)–(10). For subject JC, the coefficients w_m and w_t equal 0.58 and 0.42; $w_m = 0.45$ and $w_t = 0.55$ for subject CH; $w_m = 0.72$ and $w_t = 0.28$ for subject JH. Thus, the motion cue was mildly more reliable than the texture cue for subject JC (i.e. depth estimates based on the motion cue were less variable than estimates based on the texture cue); for subject CH, the texture cue was mildly more reliable; for subject JH, the motion cue was strongly more reliable.

The three graphs in Fig. 3 compare the optimal responses with the subjects' responses. The horizontal axis of each graph gives the values of optimal(d), the predicted average responses to texture-and-motion stimuli depicting cylinders of depth d for each of the 20 possible values of d; the vertical axis gives a subject's actual average responses (the dashed diagonal line indicates where the data would lie if the predicted and actual responses are identical). For subject JC, the correlation between the optimal responses and the actual average responses is 0.96; the correlation for subject CH is 0.95; the correlation for subject JH is 0.99. For subject JC, a linear regression in which the optimal response is the independent variable and the actual average response is the dependent variable yields a slope of 1.34 and intercept of -47.0; the slope and intercept are 1.18 and -6.68 for subject CH; the slope and intercept are 1.35 and -53.08 for subject JH. Based on these data, we conclude that the optimal cue combination model OM-U is a good model of subjects' cue combination strategies under the circumstances studied here. This is a surprising result, particularly considering the fact that the model is strongly constrained (it is linear) and it does not have any free parameters. There is a strong linear relationship between optimal responses predicted by model OM-U and subjects' actual average responses. Nonetheless, a visual inspection of Fig. 3 suggests that there is room for improvement in how well a linear model's predictions can match the subjects' actual responses, and so we consider optimal cue combination model OM-C.

Model OM-C includes the assumption that subjects are biased toward believing that cylinders are approximately equally deep as wide when considering individually either a motion cue or a texture cue, though this assumption is not used when considering both cues. As evidenced by the bias data illustrated in Fig. 2, and as was discussed above, this assumption is supported by the experimental data. Using model OM-C, the optimal response is

$$\text{optimal}(d) = w_m \langle \hat{d}_m \rangle + w_t \langle \hat{d}_t \rangle - 2 w_P d_P^* \quad (21)$$

where d_P^* is the expected value of a cylinder's depth based upon a subject's prior distribution of depth values. Using the assumption that cylinders are approximately equally deep as wide, we set d_P^* equal to the width of the cylinders (160 pixels). The linear coefficients w_m, w_t, and w_P were computed based on the variances of a subject's responses to motion-informative and texture-informative stimuli, and based on the variance of a subject's prior distribution of depth values (see Eqs. (13)–(15)).

The variance of a subject's prior distribution of depth values, denoted σ_P^2, is the only free parameter in model OM-C. The value of this parameter was set by hand so that the model's predictions closely matched each subject's responses (in the sense that the selected value

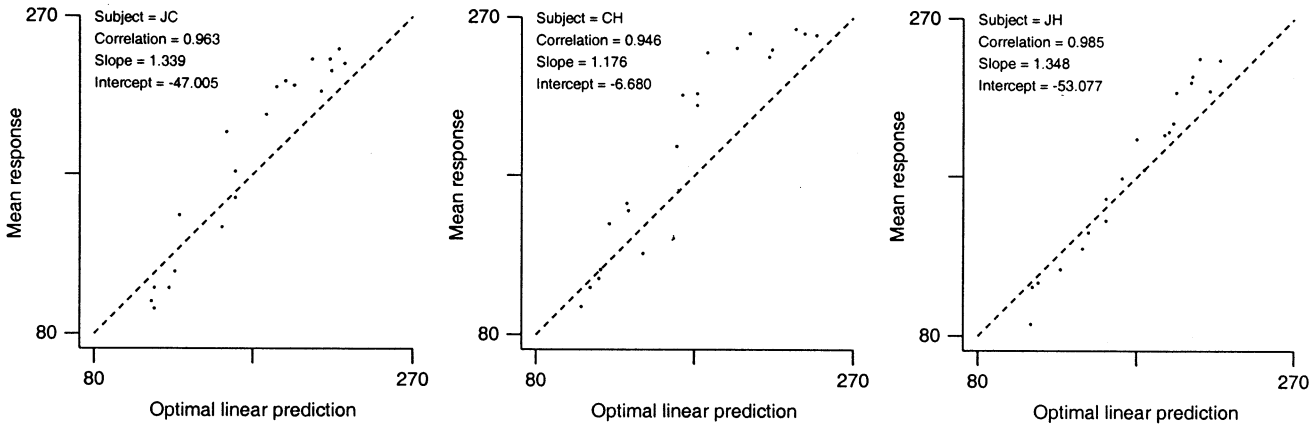

Fig. 3. The horizontal axis of each graph gives the optimal responses according to the first optimal cue combination model (OM-U) to texture-and-motion stimuli depicting cylinders of depth d for each of the twenty possible values of d; the vertical axis gives a subject's actual average responses. The dashed diagonal line indicates where the data would lie if the predicted and actual responses are identical.

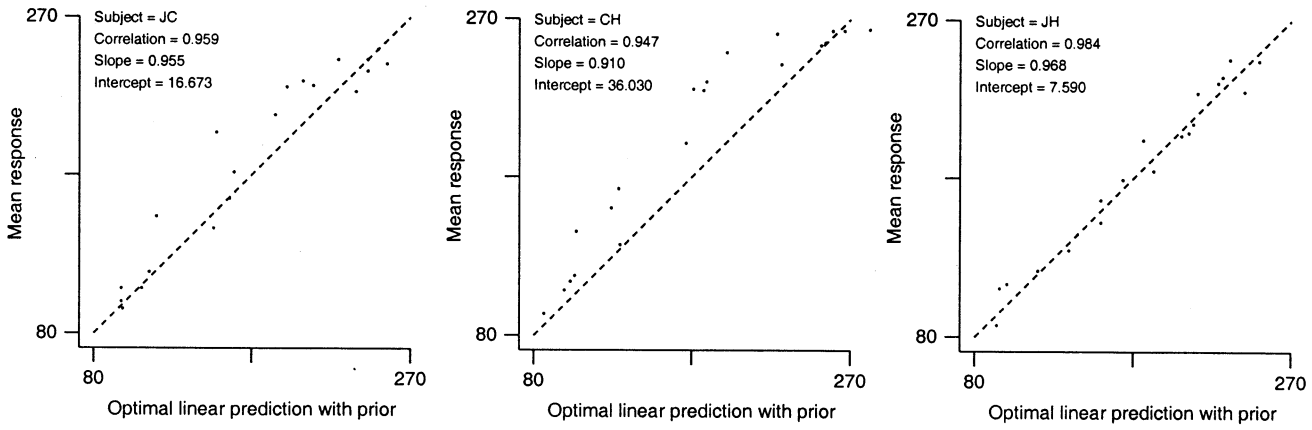

Fig. 4. The horizontal axis of each graph gives the optimal responses according to the second optimal cue combination model (OM-C) to texture-and-motion stimuli depicting cylinders of depth d for each of the 20 possible values of d; the vertical axis gives a subject's actual average responses. The dashed diagonal line indicates where the data would lie if the predicted and actual responses are identical.

cylinder of depth d. Let $\langle \hat{d}_m \rangle$ and $\langle \hat{d}_t \rangle$ denote the the model's predictions and the subject's responses). The value of the standard deviation σ_P was set to 55, 70, and 40 for subjects JC, CH, and JH, respectively. These values are sensible given the experimental data. Subject JH, for example, was most strongly biased toward assuming that cylinders are approximately equally deep as wide (see the bias data in Fig. 2), which is consistent with a prior distribution with a small variance. Subject CH was least strongly biased, which is consistent with a prior distribution with a large variance. Once the variances of the prior distributions are specified, the linear coefficients can be computed. For subject JC, the coefficients w_m, w_t, and w_P equal 0.94, 0.49, and 0.22, respectively; $w_m = 0.55$, $w_t = 0.73$, and $w_P = 0.14$ for subject CH; $w_m = 1.03$, $w_t = 0.37$, and $w_P = 0.20$ for subject JH.

The three graphs in Fig. 4 compare the optimal responses produced by optimal cue combination model OM-C with the subjects' actual average responses. For subject JC, the correlation between the optimal responses and the subject's responses is 0.96; the correlation is 0.95 for subject CH; the correlation is 0.98 for subject JH. For subject JC, a linear regression in which the optimal response is the independent variable and the actual average response is the dependent variable yields a slope of 0.96 and intercept of 16.67; the slope and intercept are 0.91 and 36.03 for subject CH; the slope and intercept are 0.97 and 7.59 for subject JH. Relative to model OM-U, model OM-C seems to provide a better fit to the experimental data as evidenced by the values of the slope and intercept for each of the three subjects.

Based on these results, we conclude that optimal cue combination model OM-C provides an outstanding fit to the experimental data. This is the case despite the fact that the model is linear, and despite the fact that it has only one free parameter. This parameter is related to the model's use of the assumption that cylinders are approximately equally deep as wide when considering individually either a motion cue or a texture cue, though this assumption is not used when considering both cues. The excellent fit between the model's predictions and the subjects' responses suggests that human observers may be making this assumption too. This conjecture is supported by the experimental data regarding the biases in subjects, responses.

In summary, we have reported the results of a depth-matching experiment in which subjects were asked to adjust the height of an ellipse until it matched the depth of a simulated cylinder defined by texture and motion cues. In one-third of the trials the shape of the cylinder was primarily given by motion information, in another one-third of the trials it was given by texture information, and in the remaining trials it was given by both sources of information. Two optimal cue combination models were described where optimality was defined in terms of Bayesian statistics. The parameter values of the models were set based on subjects' responses on trials when either the motion cue or the texture cue was informative. These models provided predictions of subjects' responses on trials when both cues were informative. The results indicate that optimal model OM-U provides a good fit to the subjects' data, and model OM-C provides an exceptional fit. Because the predictions of the optimal models closely match the experimental data, we conclude that observers' cue combination strategies are indeed statistically optimal, at least under the conditions studied here.

Acknowledgements

I thank M. Banks and I. Fine for commenting on an earlier version of this manuscript, and A. Pauls and M. Saran for help in conducting the experiments. This work was supported by NIH grant R29-MH54770.

References

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.

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

Casella, G., & Berger, R. L. (1990). *Statistical inference*. Pacific Grove, CA: Wadsworth.

Caudek, C., & Proffitt, D. R. (1993). Depth perception in motion parallax and stereokinesis. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 32–47.

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.

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.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth. In W. Epstein, & S. Rogers, *Perception of space and motior*. San Diego: Academic Press.

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.

Hearn, D., & Baker, M. P. (1997). *Computer graphics (C version)*. Upper Saddle River, NJ: Prentice Hall.

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

Johnston, E. B., Cumming, B. G., & Parker, A. J. (1993). Integration of depth modules: stereopsis and texture. *Vision Research*, *33*, 813–826.

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.

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.

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.

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.

Turner, J., Braunstein, M. L., & Andersen, G. J. (1997). The relationship between binocular disparity and motion parallax in surface detection. *Perception and Psychophysics*, *59*, 370–380.

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.

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill, & W. Richards, *Perception as Bayesian inference*. New York: Cambridge University Press.