Optimal integration of shading and binocular disparity for depth perception

Paul G. Lovell
Vision Lab, School of Psychology, University of St. Andrews, St. Andrews, UK

Marina Bloj
Bradford School of Optometry and Vision Science, University of Bradford, Bradford, UK

Julie M. Harris
Vision Lab, School of Psychology, University of St. Andrews, St. Andrews, UK

We explore the relative utility of shape from shading and binocular disparity for depth perception. Ray-traced images either featured a smooth surface illuminated from above (shading-only) or were defined by small dots (disparity-only). Observers judged which of a pair of smoothly curved convex objects had most depth. The shading cue was around half as reliable as the rich disparity information for depth discrimination. Shading- and disparity-defined cues where combined by placing dots in the stimulus image, superimposed upon the shaded surface, resulting in veridical shading and binocular disparity. Independently varying the depth delivered by each channel allowed creation of conflicting disparity-defined and shading-defined depth. We manipulated the reliability of the disparity information by adding disparity noise. As noise levels in the disparity channel were increased, perceived depths and variances shifted toward those of the now more reliable shading cue. Several different models of cue combination were applied to the data. Perceived depths and variances were well predicted by a classic maximum likelihood estimator (MLE) model of cue integration, for all but one observer. We discuss the extent to which MLE is the most parsimonious model to account for observer performance.

Keywords: 3D surface and shape perception, binocular vision, depth, shading, shape and contour


Introduction

Cues to depth and shape

There are many sources of visual information that can be used to interpret the depth and shape of objects. Some, such as binocular disparity, are reliable but can be inaccurate if the viewing distance is not known (e.g., see Brenner & van Damme, 1999; Johnston, 1991; O’Kane & Hibbard, 2010). Others, such as shading and texture, are inherently more ambiguous. For example, although we can clearly perceive shape from shading in Figure 1, the pattern of shading is determined not only by the observer–object viewing geometry, and the underlying shape of the object, but also the material properties of the object surface (its patterning) and direction and form of the light source. We know that the visual system makes assumptions about layout and lighting. For example, there is a bias toward seeing objects as convex (e.g., Langer & Bulthoff, 2001) and one that suggested that the visual system assumes light comes from nearly above (Kleffner & Ramachandran, 1992; Mamassian & Goutcher, 2001; Metzger, 1936; Sun & Perona, 1998). These assumptions allow consistent and stable perceptions of depth and shape to occur for these ambiguous cues in isolation. However, changes unrelated to the object, such as varying lighting direction, can affect perceived shape (e.g., Nefs, Koenderink et al., 2005). We also know that less ambiguous cues, like binocular disparity, can disambiguate cues such as texture (Adams & Mamassian, 2004).

Shape from disparity

Several studies have measured the reliability of disparity as a cue to shape, in particular, to obtain slant, and
demonstrated how the cue changes reliability with viewing distance. For example, Banks, Hooge, and Backus (2001) found the variance on slant from disparity estimates to increase more than 10 fold when viewing distance was changed from 40 cm to 120 cm. However, reliability also depends on the shape itself. Knill and Saunders (2003) found that slant thresholds fall by as much as a factor of 10 between a flat frontoparallel surface and one slanted at 80 degrees, but there were large interobserver differences. Shape reliability for curved convex objects is less well explored. Johnston (1991) found thresholds for judging objects with a cylindrical depth profile to be as low as 5–10% of the depth diameter, and others have also found low thresholds but often large biases in perceived shape (see also Brenner & van Damme, 1999; Hibbard & Bradshaw, 2003; Scarfe & Hibbard, 2006). In sum, it seems that for convex objects that are close enough to reach, binocular disparity can be a highly reliable depth and shape cue.

Shape from shading

Luminance gradients, which arise from shading, depend on and provide clues to the surface orientation (shape) and potentially distance between the surface and the light source and are, therefore, crucial in enabling a visual system to recover 3D information. This notion is well known and forms the basis of computational shape-from-shading algorithms. As discussed above, shading information on its own in a 2D image is an ambiguous cue to shape for the human visual system (Curran & Johnston, 1996) and additional sources of information, such as outline (Knill & Kersten, 1991), occluding boundaries (Ramachandran, 1988), surface texture (Curran & Johnston, 1996), cast shadows (Erens, Kappers et al., 1993), or prior assumptions about the probability of scene properties such as light source position (Adams, 2007; Liu & Todd, 2004; Sun & Perona, 1998) must be supplied to fully specify 3D shape.

The first formal analyses of shape from shading to come from the computer vision community (see Horn, 1975) were not so much concerned with the processes underlying how humans infer shape from shading, rather they were concerned with finding any process that might do this successfully. These early attempts were only partially successful; the correct recovery of shape required a single distant light source illuminating an object with a smooth surface viewed from some distance—essentially giving an orthographic projection of an uncomplicated surface. These conditions are rarely encountered in day-to-day life and fuller theories of shape from shading needed to account for more complex situations. Pentland (1982a, 1982b) presented a model that assumed an isotropic distribution of surface orientations, thus simplifying the task of inferring the direction of illumination and the subsequent recovery of object shape. Observer estimates of lighting direction matched those of Pentland’s model, including incorrect estimates. The variances of the model’s estimates also matched those of human observers. Todd and Mingolla (1983) presented observers with computer-rendered images of cylinders (illuminated from the side) with shiny or matte surfaces, then asked them to rate the curvature. Ratings were generally reliable, correlating strongly with actual curvature, more so for the shiny surfaces. Todd and Mingolla found that estimates of lighting direction were more reliable than curvature estimates for the same stimuli.

Mingolla and Todd (1986) favored the approach of Koenderink and van Doorn (1982), where shape from shading is based on the analysis (either perceptually or mathematically) of the relationship between the positions of local minima and maxima. Langer and Bulthoff (2001) examined depth discrimination in rendered scenes, and they found that observers could discriminate concavities from convexities reliably (>80% correct) for above-left, above-right, and diffuse lighting. They also found that performance was poorer for below the horizon illumination.

Recent studies have looked more closely into the component parts of illumination phenomena: specularities (e.g., Fleming, Torralba, & Adelson, 2004; Nefs, 2008), interreflections (Bloj, Kersten et al., 1999), lighting direction (Adams, 2007; Sun & Perona, 1996), and the
light-field characteristics (Koenderink, 2007; van Doorn, 2011). It is sufficient for the current article to accept that human observers are able to use a surface shading cue to make relative judgments of depth.

Note that there is also a binocular shading cue, which we will refer to as shading-based disparity. Because each eye views the world from a different vantage point, the relationship between object, eye, and light source is slightly different for each eye, resulting in a different pattern of shading in each eye’s view. The relative effectiveness of this binocular shading cue compared to the monocular one we have discussed above will be considered later in the paper.

Cue combination models

If several cues are available to estimate shape, a key question in understanding the underlying brain processes is how they combine and interact to deliver our perceptions. In the disambiguation of shape and shading, any increase in the knowledge of one source of information should result in the more accurate and/or reliable recovery of the other. For example, O’Shea, Banks et al. (2008) found that estimates of illumination direction become more accurate when the shape of the illuminated object was known. When only shading was available, i.e., disparity cues were removed, then estimates of the illumination direction became less accurate. Disparity has also been shown to disambiguate shape from texture information when texture could represent either a concave or convex surface (Adams & Mamassian, 2004).

A classic modeling approach that has proved successful in predicting how different cues are combined is called maximum likelihood estimation (MLE; Clark & Yuille, 1990). MLE models assume that there are independent modules that estimate depth separately and independently for each cue. The estimates are then combined by taking a weighted average, where the weights are determined by some measure of cue reliability. For sensory cue interactions between modalities, i.e., vision and touch (e.g., Ernst & Banks, 2002), models that weight cues according to their reliability provide a good prediction of human perceptual performance.

Combining disparity and shading for depth perception

While it is clear that shading in two-dimensional pictures provides a compelling sensation of depth, there is just a small body of research studying how useful shading is when combined with strong and reliable cues, like binocular disparity. Shading is generally considered a poor cousin to binocular disparity when judging the depth of nearby surfaces. The first major study in this area was performed by Bülthoff and Mallot (1988). Of particular relevance to our work is their distinction between shading-based disparity and standard point-for-point binocular disparity. While sharp edges and local features provide a reliable disparity cue, shading cues also differ for the left and right eyes. Bülthoff and Mallot found that including the shading-based disparity cue led to a stronger percept of depth than using monocular shading (where each eye sees an identical shading pattern) alone. Overall, they found that the magnitude of perceived depth decreased as individual cues were removed. We interpret this as a form of cue conflict: “removing” a cue to shape involved setting that cue to being consistent with a flat object; thus, it conflicted with the other cues to depth.1

In other studies on combining disparity and shading, it has been shown that reliability improves as more information is provided, but judgments do not necessarily become more accurate (Vuong, Domini et al., 2006). Judgments of depth may also be achieved more quickly when shading and disparity are combined (Schiller, Slocum et al., 2011; Zhang, Weiner et al., 2007). Mingolla and Todd (1986) found that observers underestimated surface curvature in non-stereo displays compared with stereo viewing. Doorschot, Kappers et al. (2001) manipulated lighting direction and interocular spacing when photographing stimuli and found that estimates of surface depth were predicted by linear combinations of the shade and disparity cue strengths.

Current study

In this paper, we conduct two experiments to determine how binocular disparity and monocular shading (each eye sees an identical shading pattern) cues are combined and to consider their relative effectiveness. A major reason for this choice of cues is that disparity is a cue that (assuming appropriate scaling by vergence or some other distance metric) can deliver a metric estimate of depth. On the other hand, shading is inherently ambiguous, requiring knowledge of the light source type and direction, before depth can be extracted. Although cue combination models have proved highly successful for a range of cue combination comparisons, one might expect them to falter here, when inherent ambiguity of one cue is very high.

Our aim is to compare several possible models of cue combination to test which the brain might be using. As well as revealing the mechanics of cue combination, our results will also inform on the extent to which the potentially ambiguous cue of shading affects shape from binocular disparity.

General methods

Using two experiments, and modeling their results, we examined the combination of binocular disparity and
shading cues in a depth discrimination task, where each cue could be explored alone (Experiment 1) or in combination/conflict (Experiment 2). Below, we outline the general methods for stimulus production and the basic procedure used for measuring depth discrimination. Further details of how we tailored specific stimuli to provide equivalent cue conflict for all observers will be described separately under Experiment 2.

**Apparatus**

Stimuli were viewed through a Wheatstone stereoscope with a total viewing distance from eye to screen of 810 mm. The stereo images were presented in 42-bit color, side by side on a calibrated Cambridge Research Systems Visage display system and CRT. Responses were recorded using a Cambridge Research Systems RB-530 response box.

**Stimuli**

Stimuli were comprised of a rendered (ray-traced) scene featuring a largely flat frontoparallel surface with a smooth (Gaussian) convexity in the center. The shape of the convex bump was defined by

\[ G(x, y) = \left( \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \right) d, \]

where \( x \) and \( y \) are the distance from the center of the scene, \( \sigma \) is the standard deviation of the Gaussian envelope, and \( d \) is the maximum depth of the convexity, measured from the plane of the screen, located 810 mm from the camera position.

The diameter of the convexity at half-height, for the “standard” stimuli depth (75 mm), was 88 mm, subtending an angle of 6.26 deg. Depth and shape in the stimuli could be defined by shading, by disparity, or by both cues (Figure 3). In Experiment 1, stimuli either contained one or the other of shading and disparity cues.

**Shading cue**

The stimuli were rendered using RADIANCE (an open-source software package validated as delivering realistic rendering; Ruppersberg & Bloj, 2006) as a uniform material object, illuminated from above and, as a consequence, featured accurately rendered attached shadows. The surface material properties were defined within RADIANCE, where the material was specified as plastic and specularity was 0.0, giving a matte surface with Lambertian reflectance. Cast shadows were not excluded; however, the geometry of the scene ensured their influence was minimal. The position of the illumination and size of the convexity were arranged so that, for \( d = 75 \) mm, the cast shadow in the rendered image was relatively small and was not clipped by the edge of the image.

Light locations were defined using the Matlab sphere function with a radius of 2876 mm and an XYZ location of [0 0 0] with a location density that results in 400 light positions placed within the intersection of two orthogonal spherical caps (see Figure 2). These were arranged so as to conform to the rule that lights should be above the horizon and in front of the illuminated scene and that lights were included only if they were within a distance of 2876 mm from an XYZ location of [575 0 2876] relative to the sphere located at [0 0 0]. This dense array of point lights was chosen so that the distribution of lighting might be manipulated easily in future studies. The planar surface (when depth was zero) was located at [0 0 0]. This resulted in a smoothly illuminated scene, which was predominantly illuminated from above, with minimal cast shadow. Shading stimuli were rendered for a cyclopean view (a view from a point exactly between the eyes), and the same cyclopean view was presented to the left and right eyes within the stereoscope. This means that the small differences in luminance gradient that provide a weak cue to depth, shading-based disparities (see Bülthoff & Mallot, 1988), were absent here. We therefore describe shading as a monocular cue to depth in this study. In the

![Figure 2. Spatial arrangement of the Gaussian convexity upon a frontoparallel surface. Illuminated from above by a diffuse light source consisting of 400 lights that fell within an area defined by two spherical caps (shown as blue planes) and within a radius of 2876 mm of a point 575 mm in front of and 2876 mm above the center of the surface. The asterisks show the camera positions and the viewing direction for the right (red) and left (blue) eyes.](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933488/ on 11/27/2018)
shading-only stimuli (Figure 3b), there were no dots added, as adding these dots with either noisy disparity or flat disparity would present a conflict to the shading cue. We chose this relatively degraded stimulus shape because it is the only shape we were able to develop that gave a reliable shading cue without delivering other shape cues, such as edges or outlines. Our purpose here was to compare binocular disparity and shading cues, so we wanted to avoid the inadvertent inclusion of other cues.

**Disparity cue**

Dots placed upon the surface of the scene simulated variations in reflectance: In other words, the object was patterned to provide consistent disparity information all over the object. Furthermore, the spatial location of the dots across the surface corresponded to the locations of equally spaced dots when placed upon the flat surface. Thus, there was no variation in texture or dot density in the final image. The disparity-only stimuli (Figure 3a) did not contain any shading information.

Our aim was to generate a surface texture of dots to deliver disparity. The dots were to be randomly located, but uniformly spaced in the image, to control for texture density cues to shape. We start with a grid of XYZ locations covering the surface of the stimulus in the 3D world, where the Z locations were determined by the depth of the stimulus bump. For each XYZ location, a small sphere was placed into the world and an image was rendered. This

---

**Figure 3.** Sample stimuli, arranged for cross-fusing (right-eye view on left and left-eye view on right). (a) Disparity-only stimuli with no shading. (b) Shade-only stimuli; note that the shading is cyclopean, and left- and right-eye images are identical. (c) The combined stimulus with monocular shading and disparity dots.
was repeated for each eye’s viewpoint and for each stimulus depth. We then recovered a corresponding set of image locations \((xy)\) for the left and right eye views, for each \(XYZ\) location in the world. Next, we defined a notional cyclopean image containing a randomized, non-overlapping set of points (2500) with uniform density. In order to establish the \(xy\) offset between the cyclopean image and each eye’s image, we calculated the differences between corresponding \(xy\) image locations. For each location in the cyclopean image, we were then able to interpolate an \(x\) and a \(y\) offset using cubic interpolation (griddata, Matlab, R2008a), utilizing the world \(XYZ\) locations and the cyclopean image \(xy\) offsets that were previously calculated. This gave us \(xy\) locations for dots in the left/right eyes’ images that deliver binocular disparity but not texture density.

Due to the left–right displacement of the cameras in the stereoscopic images and the relatively restricted field of view adopted during rendering, some dots were clipped toward the left and right edges of the images. These dots were filled-in with dots of zero disparity. This led to a small step edge along the edge of the bump. This was constant across all stereo images and did not encroach upon the center of the bump, so we felt it was acceptable in the current study.

Texture element size was not manipulated, so both the texture gradient (i.e., dot-element spacing) and dot size signaled a uniformly flat surface. This could be considered a conflicting cue as it would have indicated a flat surface while the shade and disparity cues indicated a convex surface. These confounds were unavoidable; had they been made consistent with the other cues, then our stimuli would have been supported by these texture cues as well, another potential confound.

Cue combination and conflict

In Experiment 1, we measured observers’ performance when using single cues only. In Experiment 2, we manipulated the stimuli to deliver cue combination (Figure 3c) and cue conflict. The binocular disparity of the stimulus dots was derived from the depth profile of a different depth to that represented by the shading scene. Further details of how stimuli were set up for this experiment are explained below, under Experiment 2.

Participants and inclusion criteria

Ethical approval was given by the University Teaching and Research Ethics Committee (UTREC), University of St. Andrews. Observers were recruited using the online participant recruitment system (SONA) at the School of Psychology, University of St. Andrews. Observers were compensated for their participation.

Screening. Seven observers were presented with the combined disparity and shading stimulus in the stereoscope. By rotating a dial (Cambridge Research Systems CB7), either the disparity or the shading depth could be independently adjusted between flat (0 mm) and protruding (200 mm).

All seven observers could perceive these dual-cue stimuli as a curved extrusion on a flat surface, illuminated from above. When the adjustment dial was rotated, the shading cue could be gradually removed, thus presenting a pure binocular disparity cue much like a random dot stereogram (Julesz, 1971). Observers were shown sample stimuli, while binocular disparity and shading depth were adjusted, and were asked whether they could see the convex bump. At this point, two observers were found to be unable to perceive the stereo stimuli when presented in isolation. One recalled that she had been told that she had a “lazy-eye” (amblyopia) while another simply could not perceive the disparity-only stimuli. These two observers were excluded from the study. The five remaining observers were able to perceive all versions of the stimuli; these individuals participated in the experiment. Two were undergraduates of the University of St. Andrews, and two were post-graduates and an author (PGL). Apart from the author, all observers were naive to the purpose of the experiment. Observers had a mean age of 25 years; all had corrected-to-normal vision; one wore glasses (PGL); two observers were male. The non-author observers have been anonymized.

Experiment 1

In this experiment, the aim was to compare the reliability of depth perception from the two cues, disparity and shading, independently. Details of apparatus, stimuli, and participants are given in the General methods section above.

Methods and procedure

We used a two-interval forced choice presentation. In each experimental trial, observers were presented consecutively with two stereo images and asked to indicate, by pressing a button, whether the stimulus in the first or second interval featured a deeper “bump” in the center of the screen.

Stimuli were visible for 2 s with a blank screen visible for 1 s in between, followed by another blank screen. If the observer made no response within 5 s, a prompt screen was displayed illustrating the response options (Figure 4). One stimulus (randomly in the 1st or 2nd interval) was the “standard” and was given a peak-to-base depth of 75 mm, corresponding to 23 arcmin of disparity at the peak. The
other stimulus was the test, and its depth could vary from 0 to 200 mm on any trial. Test depth was chosen using two interleaved PEST staircases, tracking the 25 and 75% points, respectively, for the “is deeper” responses (Taylor & Creelman, 1967). The starting levels for each staircase were selected so that they were likely to be at the opposite side of the point of subjective equality (PSE) for each function, ensuring that the whole range of the psychophysical function was traced by the time the staircases had converged.

At the end of each session, the staircase parameters were retained and staircases were resumed in subsequent sessions. Observers participated in at least four sessions, on separate days, each consisting of 220 randomly sequenced trials. Where the staircases had not converged, additional trials were undertaken. This was necessary for observers SQ and EB, who undertook two and one additional sessions, respectively.

In order to prevent observers from learning the standard depth and simply responding to this, a random depth offset (-10, -5, 0, +5, +10 mm) was added to both the standard and test stimuli in each trial. Where isolated cues were presented, i.e., shade or disparity only, then the isolated cue was shown in both the standard and test intervals. For cue combination, both cues were also shown in both test and standard intervals.

**Results**

We measured the response variances from the psychometric functions obtained, for individual cues and for combinations of these cues. Psychophysical functions were fit (psignifit; Wichmann & Hill, 2001a, 2001b) for each observer for shading-only and for disparity-only stimuli (Figure 5). The variance of each function was calculated by taking the difference between depth at the 50% shallower and 75% shallower levels from the fitted psychometric function.

Variances are plotted for each observer in Figure 6. For all observers, variance was higher for the shading cue than the disparity cue.

**Discussion**

When other cues to depth are controlled (i.e., texture gradient, shading-based disparity, occlusion) or at least minimized (i.e., edges and shadows), we found that responses to the disparity channel were generally more reliable than those based solely upon shading. There was an increase of 46% in the variance for shading over disparity (average across all observers). However, it would be risky to generalize from these data to other stimulus

---

**Figure 4.** Illustration of the stimulus presentation sequence and timing.

Downloaded From: https://jov.arvojournals.org/pdfsaccess.ashx?url=/data/journals/jov/933488/ on 11/27/2018
sets as variations in the lighting conditions and objects being viewed could weaken the shading cues and might decrease performance.

The next step was to perform a cue combination experiment to explore how information from the two cues was combined for depth perception. Rather than simply performing a cue combination experiment, and modeling variance, we chose to measure performance under cue conflict (so that models were challenged to predict differences in PSE, as well as in variance). In Experiment 2, we sought to carefully deliver cue conflict that was equivalent for each observer and tested which of a range of simple cue combination models could best explain the data.

### Experiment 2

Here, the aim was to explore how binocular disparity and shading cues were combined when the cues were placed in conflict (where each specified a convexity of a different peak depth). We knew from Experiment 1 that different observers showed different variance when using the disparity or shading cues alone to perform the depth task. Our first step here was therefore to design specific sets of stimuli for each observer, which had the same relative reliability of the shading and disparity cues. Our second step was to use these participant-tailored stimuli to study perception of depth during cue conflict at three different levels of relative reliability (stereo more reliable, cues roughly equated, shading more reliable; see Figure 8). These manipulations allowed several basic models of cue combination to be compared (single cue used, cue averaging, channel switching, and maximum likelihood estimation). These will be described in more detail below.

#### Methods and procedure

Details of apparatus, stimuli, and participants are given in the General methods section above.

#### Tailoring stimuli to equate for relative reliability

In order to adequately examine and model how binocular disparity and shading cues are combined, we need to vary the reliability of at least one channel while holding the other constant. The disparity channel was the most reliable channel (Figures 5 and 6) so we sought to add varying levels of disparity noise in order to manipulate the reliability of the disparity cue. It is also likely that individuals vary in their susceptibility to disparity noise, so these manipulations were tailored to each observer.

Noise was added by displacing dots by a random X offset drawn from normally distributed noise (with a mean of zero) with a standard deviation of either 0, 1.25, 2.56, or 3.86 arcmin. There were different random offsets for each dot and each eye. Two to three sessions of the depth discrimination task were administered, each with 220 trials, to estimate the response variance of each observer as a function of this disparity noise. Plotting response variance (the slope of the psychometric function) as a function of disparity noise standard deviation allowed us to select disparity noise levels that resulted in response variances that were 0.5, 1, and 3 times the response variance measured for the noiseless shading stimuli. In some instances, observers were relatively poor at discriminating the disparity stimuli with 1 arcmin or greater, so predicting appropriate noise levels would have otherwise required...
extrapolation. For these observers, the noise-balancing sessions were repeated with smaller disparity increments.

The variances were then expressed in units relative to the variance for the shading cue and plotted as a function of disparity noise level (Figure 7, red circles). These relative channel variances were then interpolated with a quadratic polynomial (polyfit with n = 2, Matlab 2008a). Where the polynomial fit was poor, i.e., variance not increasing smoothly as a function of disparity noise, then additional observation sessions were undertaken with a wider range of disparity noise levels. Finally, three levels of disparity noise were chosen, where fitted channel reliability (the variance) should be 0.5, 1.0, and 3.0 times that of the shading cue.

**Procedure**

In a series of depth discrimination trials, binocular disparity noise levels (×3) and the conflicts between the shade and disparity depths (0 mm: cue combination, ±5 mm, ±10 mm) were permuted. The direction of the depth conflicts was counterbalanced so that the shading and disparity cues were equally likely to be deeper or shallower than the other channel. In this stage of the experiment, there were a total of 17 different combinations of disparity and shading levels.

Where a cue conflict existed in a stimulus, the standard stimulus featured the conflict while the test stimulus featured no conflict (following Ernst & Banks, 2002). The presentation order of the standard and test intervals and of stimulus conditions was randomized.

**Models**

The variances and PSEs for observers were compared with the predicted variances and PSEs for various cue combination models, as detailed below. Our aim here was to test a variety of simple models, some of which would be rather unlikely. For example, we know that using only one or the other cue does not reflect what the visual system is likely to do. Rather, the point of testing such models was to see how much better the more sophisticated models could account for human perception.

*Disparity-only and shading-only.* All responses were based solely on the performance in the single-cue conditions. This simulated observers that behaved exactly in the cue-combined condition as they did in the single-cue trials, ignoring either the shade or disparity cue entirely.

*Simple-averaging model.* This simulated mandatory fusion between the channels without any weighting for reliability. Here, we simply averaged the PSEs and variances for each channel.

*Channel-switching model.* It is possible that both cues are independently available to observers and that they switch to the cue that is most reliable, ignoring the other cue entirely. In this model, the variances of the observer performance for the isolated cues were compared and PSEs and channel variances were drawn from the channel with the lower variance.

*Maximum likelihood estimation (MLE) model.* One of the most successful models for describing how visual cues combine for depth perception is the maximum likelihood estimator (MLE; Clark & Yuille, 1990). The essence of the model is that information from different cues is treated as completely independent. Information from the cues is combined optimally, so that more reliable cues are given proportionately more weight than less reliable cues. When combining cues, the weight of each channel is based on the variances of each channel:

\[
 w_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2} \quad \text{and} \quad w_d = \frac{\sigma_d^2}{\sigma_s^2 + \sigma_d^2},
\]

where \( w_s \) is the shade channel weight, \( w_d \) is the disparity channel weight, and \( \sigma_s \) and \( \sigma_d \) are the shade and disparity variances, respectively.

The predicted point of subjective equality (PSE) is based on the weighted combination of the PSE for each channel:

\[
 \mu_{sd} = w_s \mu_s + w_d \mu_d,
\]

where \( \mu_s \) and \( \mu_d \) are PSEs for the shade and disparity channels, respectively. The predicted combined-cue
channel variance ($\sigma_{sd}$) is based on the disparity ($\sigma_d$) and shade variance ($\sigma_s$):

$$\sigma_{sd} = \sqrt{\frac{\sigma_d^2 \sigma_s^2}{\sigma_d^2 + \sigma_s^2}}.$$  (4)

MLE weights the more reliable cues more strongly than the less reliable cues. We therefore expected that for low disparity noise, the MLE-predicted PSE would be close to that specified by the disparity channel, and for high disparity noise, the MLE-predicted PSE would be close to that specified by the shading channel (illustrated in Figure 8).

There was no reason to assume that there should be any bias in the single channel depth judgments because we used a two-alternative forced choice (2AFC) paradigm. However, where disparity noise levels were highest, small shifts in the PSE were sometimes seen. We did not include these single channel biases (assumed to be random) in the combination of the channels for the MLE model. A separate analysis, with the biases included, did not improve the fit between model predictions and the observed data.

**Results**

Figure 9 shows sample data for one observer, for the lowest level of disparity noise ($\times 0.5$ shading variance), for four different levels of cue conflict. The vertical dotted lines show the expected PSE if the observer were to use only the disparity cue (red line) or shading cue (blue line). Although PSEs were close to the disparity-only values, they are slightly discrepant, showing that cue combination was occurring (there is some influence of the shading cue) rather than cue vetoing (where shading would be ignored and disparity used alone).

Figure 10 provides summary variance data, showing graphs of PSEs for each of the observers (rows), with separate columns showing results for each disparity noise level. The peak depth specified by disparity is shown by the dotted icon and that for shading by the shaded icon. PSE values found from the experiment are shown by the magenta symbols, the black lines and symbols represent the model fit to be discussed later, and the error bars show thresholds obtained from the slope of the fitted psychometric functions.

Notice that when the disparity noise was low (left column), making it the most reliable channel, observer PSEs tended to fall close to the disparity cue PSE. Conversely, when the disparity noise was high (right column), observer PSEs tended to track the shading cue PSE.

In order to evaluate the fit between the predictions of the various models and the measured observer performance, the root-mean-squared deviation was calculated for the PSEs and the fitted channel variances. None of the models evaluated required any free parameters; the parameters for

---

**Figure 8.** Schematic of conflict stimuli and their combination via MLE. Left panel shows expected behavior for zero disparity noise, where the disparity cue was more reliable. Right panel shows expected behavior for large disparity noise, where the shading cue was more reliable. Lower graphs show the MLE model prediction: The point of subjective equality (PSE) shifted toward the more reliable cue, when disparity noise was added.
each model were based on the response characteristics of each observer to each channel in isolation; consequently, there was no need to use the Akaike’s information criterion (Akaike, 1974). Instead, we compared the overall error for predictions of PSEs and response variances. Root-mean-squared deviations (RMSDs) are the usual metric for such comparisons, and we used them here.

Variance values for the model outputs tended to be larger than the PSE values, but the fits for both are equally important when evaluating the performance of the models. Consequently, the RMSD values were normalized independently for variance and PSE (by dividing by their standard deviations) before they were added together. Figure 11 (left panel) shows the total RMSD values for each model and each observer. Overall, the MLE model had the lowest RMSD value, i.e., it was the best fitting model. The MLE-predicted PSEs and variances are shown within Figure 10 as black circles and vertical bars.

The difference in the RMSD values for the three best performing models were relatively small, so closer inspection of the pattern of model predictions for PSEs was required to distinguish between them.

To test the simple average model, we explored the pattern of PSEs delivered by our observers. The simple average model predicts that all PSE values should be 75 mm, matching the average depth of the cue conflict stimuli. If this were the case for our observers, then we would expect a very narrow spread of estimated PSEs with a mean of 75 mm. Visual inspection of the magenta circles in Figure 10 suggests that this is not the case. In order to evaluate this possibility statistically, we examined the RMSD residuals found from fitting the simple average model to the observer PSEs. Any evidence of bimodality in the residuals would suggest that the observers are not simply averaging the cues, rather the observers’ PSEs differ from the 75-mm average. To test this possibility, we administered a statistical test for unimodality (the null hypotheses) to the PSE values, namely, Hartigan’s dip test (Hartigan & Hartigan, 1985; Hartigan, 1985). The residuals for the larger conflicts were bimodal (dip = 0.1028, \( p = 0.016 \)). This bimodality is illustrated in the lower inset histogram in Figure 11. The simple average model would not predict any bimodality in the PSE values; hence, we can reject this model.

The switch to minimum variance model predicts a bimodal distribution of PSE values for cue conflict stimuli. Observers will either respond to the shade or disparity cues. If this were the case, then we would expect a bimodal distribution to be observed in the PSE values themselves. To test this possibility we again administered Hartigan’s dip test, this time to the observer PSEs. In order to maximize the likelihood of rejecting the null hypothesis, that is of finding bimodality, we normalized the cue conflicts, so that the PSEs for the smaller conflicts did not mask any bimodality in the overall distribution. There was no evidence of bimodality in the distribution of normalized PSEs for any of the observers (greatest dip value = 0.1135, \( p = 0.116 \)); consequently, we can reject the switching model. Thus, MLE is the only remaining plausible model with low RMS error, as shown in Figure 11.

**Discussion**

Four of the five observers shifted their reliance from disparity toward shading as a function of stereo channel reliability (Figure 10). Further, the shift in the PSE for cue conflict stimuli was best predicted by the MLE model of cue combination, as shown in Figure 11. The variances of responses to the combined cue stimuli were also best predicted by the MLE model. RMSDs for the switching model increased by 35% on average for each observer relative to the MLE levels. However, the RMSD provides a relatively coarse estimate of model suitability that ignores the distribution of the residuals. Ideally, the residuals should be Gaussian with a mean of zero and

![Figure 9](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933488/) The four psychophysical functions represent fits for a single observer (MH) under varying cue conflict sizes and directions. Dotted and shaded icons represent the depth of the disparity and shade cues, respectively. Disparity noise was set to 0.5\( \times \) shade variance. Estimated PSEs for observer responses tracked the depth of the disparity cue. The data for this observer are summarized further in Figure 10 (top left).
Figure 10. Observer PSE (magenta circles) and variances (magenta lines) as a function of cue conflict. Cue conflict stimulus depths are indicated by shaded (shading cue) and dotted (disparity cue) icons. Results of the MLE model are shown by black lines (predicted variance) and symbols (predicted PSE).

minimal spread. Figure 11 (right inset) shows the residuals for predictions of the PSE for the three best performing models. It is clear that the MLE model had the most tightly packed histogram, while the spread was larger for the switching model. For the simple average model, the spread was almost bimodal. For the simple average model, all the predictions would fall upon the horizontal dashed lines within Figure 11 (left) so the spread of the residuals reflects the fact that rather than following the dashed lines they tend to span the line. For observer MB, the simple average model does, albeit with only a marginal decrease in mean residuals, provide the best fit. Parsimony dictates that we should accept the slightly weaker MLE fits for MB rather than unnecessarily postulating multiple models.

A fifth observer, SQ, showed patterns of performance that were not well predicted by the MLE model, nor were they predicted by any other model we tested. A closer examination of SQ’s performance levels reveals that, in the initial noise-balancing trials, only a relatively small amount of disparity noise was required to make stereo
channel variances three times larger than the measured shade channel variances—around 3 arcmin compared to nearly 6 arcmin for the other observers. Despite the relatively small amount of disparity noise, the variances recorded were the largest. SQ’s data may be understood if we accept that she was relatively poor at detecting the disparity cue when it was presented in isolation. In fact, she could see the disparity cue more reliably when it was presented with the shading cue. When both cues were present, her discrimination variances were much reduced and the relatively small levels of added disparity noise were then insufficient to cause a shift toward the shading channel.

To sum up, our data demonstrated that the two cues of shading and binocular disparity did combine to deliver a perceived shape. The model that most adequately described the data was the MLE model.

**General discussion**

**Summary of findings**

In this study, we used objects whose outline did not change with depth, so that the relative contributions of shading and disparity to shape perception could be studied independently of other cues. We used realistically rendered shading, consistent with a mixture of diffuse and directional lighting, designed to favor biases known to be inherent to the human visual system’s use of shading (e.g., directional light was from above, and all objects were convex). Shading was combined with dense binocular disparity information, delivered by dots that did not contribute to the shading profile (they were fixed at a constant luminance, whatever the object shape, and did not carry any consistent shape-from-texture information). We showed that disparity provides a more reliable cue to depth than shading, yet when disparity was compromised by noise, the shading cue could be reliable and consistent.

We used a series of cue conflict conditions, where shading and disparity delivered different peak depth information, and measured PSEs and variances of human performance in a task asking which of two convex objects had the greater depth. Human data were compared with a number of potential models of cue combination. We expected that the simplest models would be inadequate, but the key question was whether a maximum likelihood estimator (MLE) model, which posits independent and optimal combination of cues, was adequate to explain our data.

For all but one observer, MLE predicted PSEs and response variances very well for cue-conflicting stimuli.

**Figure 11.** (Left) Total normalized residuals of each model’s fit of PSEs and variances. The RMSD values for MLE tended to be lowest, however only slightly so. (Right) The pattern of the PSE residuals for the switching and simple average models suggests a poorer fit than the MLE model. The MLE residuals are Gaussian with a mean of zero, while the others have a wide spread (switching) or a bimodal distribution for the largest cue conflicts (simple average).
Other models such as channel switching resulted in poorer predictions of either the PSEs or the response variances.

Our data demonstrate that the visual system used the two cues as if they deliver totally independent estimates of depth. We do not know whether our observers have conscious access to each of these two channels, i.e., whether observers could have at will responded to the disparity or shading cues individually had they been instructed to. Pilot studies did reveal that the stimuli had to be quite carefully constructed in order to achieve a level where the shading and depth cues combined seamlessly and did not obviously break into layers when a conflict was introduced. Therefore, the current experiment demonstrates optimal combination of information from two sensory cues, but it does not necessary follow that this fusion was mandatory (Hillis, Ernst et al., 2002) with the consequence that access to the individual cues was lost. This issue is beyond the scope of the current manuscript.

The presentation order of noise levels in the cue-conflict stimuli was randomized from trial to trial, so our results must reflect “dynamic cue weighting” (Ernst & Banks, 2002). The random presentation order meant that our observers were not trained over a session to ignore stereo because it had become less reliable. Instead, they were immediately shifting their weighting, trial by trial, toward the more reliable cue. Ernst and Banks (2002) called this dynamic reweighting and suggested a potential neural substrate for online cue reweighting, where populations of neurons independently represent the depths in each channel. Taking the product of these populations facilitates the computation of the weighted mean value. Our findings require that cues are combined via dynamic reweighting.

In the sections below, we review how our results fit into the current literature, discussing the importance of careful choice of stimuli and what others have found when comparing shading and disparity cues in cue combination or cue conflict experiments.

Choice of stimuli and cue independence

In any study of cue combination, one considers a number of cues (usually two) and makes manipulations that reveal how they might be combined. For such a study, the stimulus must be very carefully constructed. Here, for example, we used only monocular shading. The left- and right-eye views contained an identical shading pattern, so that the shading itself delivered zero binocular disparity across the whole image. Thus, only shading specified depth and shape. This is clearly an unnatural situation. Because the two eyes view the world from different locations, the shading patterns at the two eyes are different; hence, the stimulus contains shading-based disparity first suggested by Mach (1866) and studied in detail by Arndt, Mallot et al. (1995) and Büthoff and Mallot (1988). However, shading-based disparity cannot be manipulated independently of monocular shading, because the former relies on the latter. Hence, we separated the two cues by only using monocular shading as our shading cue. We felt justified in making these manipulations because shading-based disparity cannot be used to obtain local variations in depth that determine shape (Hill & Bruce, 1993; Mallot, 1997; Mallot, Arndt et al., 1996).

We also took care to think about how the dots defining disparity might impinge upon the shading cue. Recent work (Sun & Schofield, 2011) suggests that shape from shading can be disrupted by low-frequency texture information but is less affected if the texture is of a much higher spatial frequency than the underlying shading pattern. Norman, Todd et al. (2004), found worse performance for shape discrimination when texture and shading were used to judge shape than when shading was presented alone. Disparity was applied to our stimulus via small densely packed dots that covered approximately 30% of the image. Dot luminance was set to be equivalent to variations in surface reflectance. Dots did not carry any consistent shape from texture information. In pilot work, we found that if the disparity-carrying dots were too few, the percept tended to break into a pair of transparent surfaces, one carried by disparity and the other by shading.

Our stimuli were designed to avoid the impact of other cues to depth and shape. For example, our smooth bumps (Figure 3) were designed so that there was no useful object contour, which has been shown to be a strong cue to depth and shape (e.g., Koenderink, 1984; Koenderink & van Doorn, 1982; Malik, 1987; Singh & Hoffman, 1999). Contour disparity, caused by each eye seeing a slightly different contour due to its different viewpoint, is also a useful cue to depth (Nefs, 2008) and was absent from our stimuli. We removed texture cues: The shaded object did not have texture itself, and texture information provided by the superimposed disparity dot pattern did not warp with the shape: Instead, it specified a flat surface.

Finally, we used a carefully considered choice of lighting and shape so that the ambiguous shape cue was as stable as possible. Shape from shading is an inherently ambiguous cue: The shading pattern changes not only with object shape but also changes with variations in the object viewpoint and changes in lighting direction (Horn, 1975, 1990; Pentland, 1989). We also know that changes in light direction can result in changes in perceived shape (Christou & Koenderink, 1997; Koenderink, van Doorn et al., 1996; Nefs et al., 2005). Presumably to combat some of this ambiguity, the visual system appears to operate several strong “priors” that bias observer judgments of shape from shading. In particular, there is a bias toward perceiving objects as convex (Hill & Bruce, 1993; Langer & Büthoff, 2001), and the visual system assumes that light comes from above (Ramachandran, 1988; Todd & Mingolla, 1983). Indeed, one study has suggested that, for shape judgments of concave objects, observers can...
behave as if they use only priors, and there is no influence of the visual input at all (Liu & Todd, 2004). There is also evidence that the convexity prior can be modified (Champion & Adams, 2007). Here, we were careful to set directional lighting to be from above, from a consistent direction on all trials and to only use convex objects. No observers reported any kind of instability in perceived object shape and our highly reliable results suggest that, for our stimulus arrangement, shading was only around 30% less reliable than binocular disparity.

Other studies on binocular disparity and shading cue combination

That shading and binocular disparity interact for shape perception is well known (e.g., Wright & Ledgeway, 2004). It has been suggested that disparity might be more important for judgments related specifically to shape and shading for judgments related to curvedness (Tittle et al., 1998).

Different studies on how shape and disparity information might be combined for depth and shape have found different results and come to different conclusions. This may be because researchers have made different choices about stimulus design. For example, Vuong et al. (2006) found complex interactions between disparity and shading cues, which could not be modeled using MLE or other simple models. They suggested something complex and non-linear was going on in how the brain combined cues. However, their experiment used shading-based disparity rather than monocular shading, so the study is actually testing the interaction of shading-based disparity, fine-scale disparity, and monocular shading. The interpretation may also have been further complicated by the presence of a visible outline cue and the use of a very sparse dot pattern to define fine-scale binocular disparity.

Bülthoff and Mallot (1988) found that where outline information was available, it was relied upon for shape judgments more than shading and disparity. Their message was that adding in additional cues delivered greater perceived depth. However, they did not model or collect data for the most powerful conditions, those where two cues deliver conflicting visual information.

Different kinds of results can also be obtained when a depth setting task is used rather than 2AFC psychophysics. Doorschot et al. (2001) used a surface normal setting task, with several combinations of disparity-defined shape and lighting direction. Their chosen objects were complex: photographs of statues of a human torso. An analysis of the causes of variance in the setting data suggested that shading and disparity cues were almost linearly combined. We do not know whether this different conclusion arose due to the very different task employed, the richer cues available from the complex object, or the fact that these authors specifically avoided using a cue conflict paradigm involving the necessary discrimination task. In future work, we intend to follow up these issues in more detail.

Summary and conclusion

For a realistically rendered shaded object, the shading cue can provide a reliable signal to the depth and shape of the object, though for the object we used, shading was almost twice as noisy as binocular disparity. On the whole, our data were very well fit by a standard MLE model, giving us no reason to require a more complex model to account for performance. Hence, despite the huge potential ambiguity associated with shading, when the conditions are right, shading can be highly reliable and combine optimally with disparity for shape perception.

Acknowledgments

We thank our observers for taking part in the experiments. This research was supported by the Engineering and Physical Sciences Research Council—UK (EPSRC) via Grant EP/G038708/1 to JMH and EP/G038597/1 to MB.

Commercial relationships: none.
Corresponding author: Julie M. Harris.
Email: julie.harris@st-andrews.ac.uk.
Address: Vision Lab, School of Psychology, University of St. Andrews, South St., St. Andrews, KY16 9JP, UK.

Footnote

An important point to note, however, is that although shading-based disparity and point-for-point disparity can be separated in the stimulus and described as having “separate” effects, we do not mean to infer that separate disparity-based brain mechanisms are required for the processing of each.

References


