Estimating changes in lighting direction in binocularly viewed three-dimensional scenes

Holly E. Gerhard
Department of Psychology, New York University, New York, NY, USA
Laurence T. Maloney
Department of Psychology, New York University, New York, NY, USA & Center for Neural Science, New York University, New York, NY, USA

We examine human ability to detect changes in scene lighting. Thirteen observers viewed three-dimensional rendered scenes stereoscopically. Each scene consisted of a randomly generated three-dimensional "Gaussian bump" surface rendered under a combination of collimated and diffuse light sources. During each trial, the collimated source underwent a small, quick change of position in one of four directions. The observer’s task was to classify the direction of the lighting change. All observers were above chance in performing the task. We developed a model that combined two sources of information, a shape map and a shading map, to predict lighting change direction. We used this model to predict patterns of errors both across observers and across scenes differing in shape. We found that errors in estimating lighting direction were primarily the result of errors in representing surface shape. We characterized the surface features that affected performance in the classification task.

Keywords: illumination perception, 3D perception, shape perception

Introduction

In everyday scenes, the spectral composition and intensity of light arriving at each point within a scene can vary markedly with direction. The light emitted by a surface toward the eye can then depend on three factors: (1) surface material properties including color and lightness, (2) scene lighting, and (3) the location and orientation of the surface within the scene. Recent research has focused on human ability to estimate surface properties despite variations in illumination and surface location and orientation (see Maloney, Gerhard, Boyaci, & Doerschner, in press for a review). There is a second literature on estimating shape (including location and local surface orientation) when illumination and surface properties vary (Khang, Koenderink, & Kappers, 2007; Nefs, Koenderink, & Kappers, 2005, 2006).

Most recently, researchers have examined how accurately human observers estimate the flow of light within the scene, the light field (Gershun, 1936/1939). Koenderink, Pont, van Doorn, Kappers, and Todd (2007), for example, asked observers to adjust the shading on spherical probes embedded in scenes to match the local light field. The pattern of shading on Lambertian (matte) reference objects present in the scenes provided the information about the light field. They found that observers could effectively interpolate the light field at locations within the scene distant from any reference object.

Gerhard and Maloney (2010) examined human ability to discriminate transformations of the pattern of scene luminance induced by altering the direction of a collimated1 light source from non-light transformations that shared the same local scene statistics. They found that observers were accurate in discriminating light and non-light transformations and that they were more sensitive to a concomitant change in the albedo of a surface patch under a light transformation than under a non-light transformation. In Movie 1 and Figure 1, we illustrate the pattern of changes in luminance induced on a Lambertian sphere by changing the direction to a collimated source to the left, right, up, and down. In this article, we examine human ability to judge such changes in light source direction and develop a model of human performance.

After inspecting Movie 1, the reader likely believes that estimating the direction of movement of a collimated source is very easy. However, there is a well-known ambiguity inherent in the estimation process illustrated in Figure 2. As an inverse optics problem, inferring the direction to the light source requires both a shading map and information about the shape of the object, the shape map. The shape map is needed to disambiguate the shading map as illustrated in Figure 2. The upper image in Figure 2 is the shading map of a corrugated surface varying sinusoidally in depth. The image could have been...
generated by either of the two surfaces shown below, one that is a reversal in depth of the other, where concavities have flipped to convexities, and vice versa. The associated estimates of light direction are opposite: in one case, the light comes from the right, and in the other, from the left.

The reader’s likely success in interpreting Movie 1 is the result of a strong bias to interpret the surface shown as convex. Indeed, we clearly have a strong prior for convexity when viewing bounded shapes and faces (Hill & Bruce, 1993, 1994; Johnston, Hill, & Carman, 1992). The shading map does not contain enough information to estimate the direction of movement of the collimated source. Here, however, the viewer likely “guesses” the correct shape map.

In past work, when monocular views of objects are presented, observers can estimate the direction to the light source with high accuracy when shape information is available, e.g., from a prior. Monocular images like the one in Figure 1 are interpreted as pictorial representations of convex three-dimensional objects resulting in high accuracy in light direction estimation for monocular views of spherical objects (Pont & Koenderink, 2007). When several monocular images of a convex object from various viewpoints are displayed supporting structure from motion, observers are similarly highly accurate in estimating both the azimuth and elevation to a light source, particularly for collimated light sources (Khang, Koenderink, & Kappers, 2006).

With some monocular images, observers cannot estimate shape veridically. In particular, rough surfaces

---

**Movie 1.** Changes in direction of illumination. We illustrate how changes in direction to a collimated light source induce changes in luminance on a lightly textured Lambertian sphere. Over the course of the movie, the collimated source, initially aligned with the observer’s line of sight, rotates left, right, up, or down by 45° in elevation.

**Figure 1.** Changes of direction of illumination. The five spheres are taken from Movie 1. Each is Lambertian (matte), and each is rendered under a collimated light source. They illustrate the changes in the pattern of luminance induced by changes in the direction to the collimated source. The collimated source illuminating the central sphere is placed along the observer’s line of sight. The spheres to the left, right, above, and below illustrate the pattern of luminance resulting when the collimated source has rotated 45° in elevation left, right, up, or down, respectively.

**Figure 2.** The shading ambiguity. The 2D shading image shown at top could have been generated by either of two sinusoidally corrugated surfaces, which are depth reversals of each other. The left surface, when combined with the shading image, is consistent with a light source on the left, while the right surface, which is opposite in depth, is consistent with a light source having the opposite azimuth, coming from the right. The consistent position of the light source is shown with each interpretation.
viewed at the mesoscale lead to errors in light direction estimation. For example, Koenderink, van Doorn, and Pont (2004) presented subjects with monocular images of shaded Gaussian surfaces and found that observers randomly committed 180° errors in estimating the azimuth of the light source. Even for familiar objects and surfaces, such as leaves, orange peels, or crumpled paper, observers commit 180° errors when the surfaces are viewed as textures rather than complete objects (Koenderink, van Doorn, Kappers, te Pas, & Pont, 2003).

In a previous study mentioned above, we presented observers with binocular views of three-dimensional scenes, where only stereo disparity disambiguated concave from convex interpretations of the objects in the scene. Observers utilized the stereo information to estimate rapid, small changes in light source direction and performed near ceiling in estimating the direction of change (Gerhard & Maloney, 2010).

There are two goals to the current work.

First, we expand on our previous psychophysical study by rendering richer, more variable stimuli (referred to as “scenes”) and testing human performance on lighting direction estimation over a variety of three-dimensional scenes whose shapes are ambiguous when viewed monocularly. In order to perform well in the experimental task, observers must combine shape and shading information.

For some scenes, we will find that observers can more readily judge changes in light direction in the horizontal direction than in the vertical, and for other scenes, we will find the opposite pattern. We will also look at how such anisotropies depend on the surface features of individual scenes.

Second, we develop a model that combines the shape and shading maps of a scene to estimate lighting direction, and we match it to human performance by introducing “noise” in shape and shading estimates. We evaluate the model by evaluating its ability to produce the exact pattern of correct judgments and errors in each observer’s data and by its ability to reproduce the anisotropies found in human data.

Experiment

Introduction

We evaluated the human visual system’s ability to rapidly estimate change in the direction of lighting in simple scenes2 illuminated by a combination of collimated (directional) light source and a diffuse (non-directional) source. The collimated light source moved in one of four possible directions (left, right, up, down) during each trial, and observers classified the direction of movement by responding “left,” “right,” “up,” or “down” (Figure 3).

Figure 3. Schematic of the task. The stimuli were hilly, textured landscapes. A collimated light source rotated through a 10° arc over the scene in one of four cardinal directions, and the observer’s task was to indicate the motion direction. Observers viewed the stimuli from a bird’s-eye view in a stereoscope. Light direction elevation was constrained to avoid cast shadows. (Inset) Projected light paths in the image plane are shown schematically in the four directions. They are arranged randomly to emphasize that the paths were randomly located but always parallel to the cardinal axes.

Methods

Stimuli

Coordinate system

In the following descriptions, we use two coordinate systems (Figure 4). The first is a right-handed Cartesian coordinate system \( p = (x, y, z) \). The \( xy \)-plane is frontal-parallel at 60 cm from the viewer, and positive \( z \) indicates coordinates nearer to the viewer than the \( xy \)-plane. Positive \( x \) runs horizontally to the viewer’s right and positive \( y \) points toward the ceiling of the room. The origin is the extension from the cyclopean point out to 60 cm in front of the viewer. Stimuli were centered on the origin. The second coordinate system in Figure 4 is spherical, \( p = (\psi, \varphi, \rho) \) with azimuth \( \psi \) and elevation \( \varphi \) analogous to latitude and longitude on an imaginary terrestrial sphere and \( \rho \) equal to distance from the Cartesian origin. The imaginary sphere is centered on the origin, and the observer’s line of sight passes through what would be the North Pole and the axis of rotation of the terrestrial sphere. We will use azimuth and elevation \( (\psi, \varphi) \) to denote directions with respect to the origin. We report azimuth and elevation values using a superscript degree symbol, e.g., 5° while we report “degrees of visual angle” with respect to the observer as “DVA.”
in a fronto-parallel plane. The spherical with coordinates azimuth passing midway between the observer Gaussians. Heights were distributed uniformly on the.

Figure 4. Coordinate systems. We use two coordinate systems. The first is a Cartesian coordinate system with the x- and y-axes in a fronto-parallel plane. The z-axis is along the line of sight passing midway between the observer’s eyes. The second is spherical with coordinates azimuth \( \psi \), elevation \( \phi \), and radius \( \rho \).

Scenes

Three hundred scenes were generated. Each consisted of 30 isotropic Gaussians placed in a 5 DVA \( \times \) 5 DVA square (the display area) within the xy-plane. Each Gaussian had standard deviation of 0.5 DVA. To determine the position \( \mu_i = (\mu_i^x, \mu_i^y) \), \( i = 1, \ldots, 30 \) of the \( i \)th Gaussian centered on the origin, we first selected a uniformly distributed location. After assigning each Gaussian a position, we then checked that all were separated by at least 1 standard deviation, and those that were too close together were randomly jittered in position iteratively until the separation criterion was met. We then assigned heights \( h_i \), \( i = 1, \ldots, 30 \), to the 30 isotropic Gaussians. Heights were distributed uniformly on the union of the two intervals [−1.5, −0.5] and [0.5, 1.5].3 After the positions and heights were determined, the 30 Gaussian functions were summed in \( z \) to form a Gaussian “bump” surface patch (an example is shown in Figure 3) with depth \( z(v) \) at any point \( v = (x, y) \) within the display area:

\[
z(v) = \sum_{i=1}^{30} h_i \frac{1}{(2\pi)^{1/2}(\sum_{j} \mu_j)^{1/2}} e^{-\frac{1}{2}(v-\mu_i) \cdot (v-\mu_i)},
\]

(1)

where

\[
\sum = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},
\]

(2)

and \( \sigma = 0.5 \) DVA. The \( z \) range was then normalized to be 2.625 cm, and the patch finally translated such that all \( z \)-values were equal to or greater than 0. All scenes therefore had the same fixed \( z \) range of 2.625 cm with the furthest point from the observer embedded in the xy-plane.

Light paths

Four light paths of a collimated source were generated for each of the 300 scenes. The light paths were always in the positive \( z \) octants, or in other words, above the scene, effectively at infinity behind the observer. The directions were fixed so that the projections of the light paths in the xy-plane would correspond to \( +x \) (right), \( -x \) (left), \( +y \) (up), and \( -y \) (down). Each light path consisted of 16 equally spaced positions along an arc of constant radius from the origin.

The collimated source can be envisioned as a fixed point on the celestial sphere centered at the origin, and motion of the collimated source can be thought of as rotations of the celestial sphere, about either the \( y \)-axis (right or left motion) or the \( x \)-axis (up or down motion).

The initial position in spherical coordinates \( p_i = (\psi_i, \varphi_i, \rho) \) of each light path was randomized; initial azimuth, \( \psi_i \), was uniformly distributed on \([0^\circ, 360^\circ]\), initial elevation, \( \varphi_i \), was uniformly distributed on \([70^\circ, 90^\circ]\), and \( \rho \) was fixed to be large and effectively infinite.

We wish to emphasize the importance of randomizing the initial positions of each light path; if every trial started with the same initial position, observers would not need to track the lighting direction over the course of the trial because the final position would directly indicate the change in direction (e.g., Figure 1). The randomization effectively precluded using the final position as a cue: some “down” trials could end with a final frame where the light comes from the right or above or even diagonally up and to the right. For the same reason, the initial frame provides nearly no information as to which response is correct.

After selecting an initial position \( p_i \), each of the 15 remaining positions along the trajectory of the light path was computed by rotating \( p_i = (x_i, y_i, z_i) \) about either the \( x \)-axis (up or down motion) or \( y \)-axis (right or left motion) in steps of \((10/15)^\circ\). This procedure is equivalent to the following: the collimated source can be envisioned as a fixed point of light on the celestial sphere centered on the origin, and its rotation above the scene occurs as a consequence of the rotation of the celestial sphere.

Rendering

Each scene was tessellated with 4,802 right triangles, each with base and height equal to 0.07 DVA. The value of the albedo of each triangle was selected uniformly from the interval \([0.4, 0.8]\). This albedo variation was necessary to provide reliable disparity information so that the scene would not appear to move during the light source rotation.

After a random scene had been generated following the above procedure, four random light paths were selected, and the angle between each tessellation triangle’s surface normal and the 64 light positions (4 paths \( \times \) 16 frames/path) was computed. If any of these angles was outside the
range \([-90^\circ, 90^\circ]\), a new random scene was generated, and the process was repeated until a set of scenes satisfying this condition for each of four light paths had been generated. This ensured that every triangle in each scene was illuminated by the collimated light source (and thereby insuring there were no shadows cast by the collimated source).

Orthographic projections of each triangle were computed for the viewpoints of the left and right eyes assuming an interocular distance of 6 cm, which was sufficient for all the observers to view the stimuli in depth comfortably and with relative depth accuracy, as confirmed in a stereo screening test described in the Observers section.

Luminance computation

The luminance of each of the 4,802 triangles was computed following Lambert’s Law (Haralick & Shapiro, 1993, pp. 2–7). The luminance of the \(j^{th}\) triangle with albedo \(a\) on frame \(t\) was determined by

\[
L^j(t) = (L_c \cos \theta_j(t) + L_d) a^j, \tag{3}
\]

where \(L_c\) is the intensity of the collimated light source, \(L_d\) is the intensity of the diffuse light source, and \(\theta_j(t)\) is the angle between the \(j^{th}\) triangle’s surface normal and the direction of the collimated light source in frame \(t\). The intensity of a collimated source is defined as the luminance of a Lambertian surface with albedo equal to 1 positioned so that its surface normal is parallel to the light source’s direction. In these terms, our collimated source’s intensity was 25 cd/m\(^2\). We set the diffuse source’s intensity at one-quarter of the collimated source’s intensity to simulate a simple “sun and sky” environment. All trial frames were rendered beforehand and displayed as 8-bit portable network graphics images.

Trials

Trials were 16 frames each. Before every trial, there was a 250-ms intertrial interval during which the screen was black. Then, the first frame was displayed for 750 ms, so that observers could fuse the stereo pair and acquaint themselves with the scene geometry and initial lighting direction. During the final 15 frames, the light source rotated through a 10\(^\circ\) arc over the scene. Each of the final 15 frames was displayed for 50 ms, for a total trial duration of 1500 ms, following which the stimulus was extinguished and the screen turned black. The experiment software did not accept subject responses until the stimulus had been extinguished. Following the recording of 1 of the 4 allowed response keys, the 250-ms intertrial interval began.

Apparatus and software

Left and right images were presented to the corresponding eyes on two 30-in. Dell UltraSharp 3008WFP LCDs with 2560 × 1600 pixel resolution and 60-Hz refresh rates. The monitors were placed to the observer’s left and right. Two small mirrors were placed in front of the observer’s eyes to reflect the images separately from each monitor into the corresponding eye. Luminance measurements from a Photo Research PR-655 spectroradiometer were taken at five points on each LCD: the center of the stimulus location, and at 2.5 DVA right, left, up, or down with respect to the center of the stimulus location. Each measurement was taken by holding the midpoint of the spectroradiometer’s lens constant while pivoting the spectroradiometer housing so the lens’ view was centered on the region of interest. The spectroradiometer’s lens was placed at the observer’s eye position in front of the mirrors. Spatial homogeneity and directional independence of the luminance signals over the portion of the screens used for stimulus display were confirmed. Separate lookup tables for each LCD were created using luminance measurements taken at each display’s center to correct for non-linearities in the liquid crystal responses and to equalize display values for the two LCDs. The maximum luminance achievable on each display was set to 67 cd/m\(^2\).

The apparatus was housed in a dedicated room against a back wall of length 192 cm. All walls of the room, the table supporting the displays, and the casings around the displays were covered with black flocked paper (Edmund Scientific, Tonawanda, NY) to absorb stray light. During the experiment, the lights were turned off and the door was closed. Only the stimuli were visible to the observer, who viewed them from a head and chin rest secured on the table supporting the displays. The casings of the displays and other features of the room were hidden from view by a black flocked tunnel connecting the chin rest to the mirrors. Additional light baffles were placed on both sides of the observer’s head to prevent the light from the LCD monitors from reaching the observer’s eyes directly. The optical distance from mirror to the corresponding LCD monitor was 60 cm, and stimuli were rendered to be 60 cm in front of the mirrors to minimize conflict between binocular disparity and accommodation. A schematic of the stereoscope is shown in Figure 5.

The experimental software was written in the MATLAB programming language using the Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997). The computer was a Dell Precision T7400 running Ubuntu LINUX 8.04 with an NVIDIA GeForce 9800 GT dual-DVI graphics card.

Procedure

Observers were instructed that they would be judging the direction of lighting motion. Prior to viewing any stimuli, observers were told that they would be viewing a three-dimensional landscape. They were told that the trials would be 1.5 s in duration, and that for the first half of the trial no change would occur, but that during the final 750 ms of the trial, the out-of-view point light source would move through a 10\(^\circ\) arc over the scene, and they
procedure involved making judgments of 20 scenes containing 8 four-sided hollow pyramids, each of which could be convex or concave. The scenes were rendered for binocular viewing and presented in the same stereoscope as was used for the experiment. In each scene, there were always either more concave or more convex pyramids, and the task was to indicate whether there were more concave or convex pyramids present. All observers completed this task quickly and performed perfectly.

All observers were reimbursed $10/h for their time except author HEG. All observers completed the experiment in one session, lasting about an hour on average.

Results

The same 1,200 pre-rendered motion trials were classified by all 13 observers. Summary results of their classification performance are reported here across observers, motion directions, and test scenes. Additional analyses of their performance based on the ideal observer model we develop below are reported in the Model section.

Each of the observer’s data falls into a $4 \times 4$ confusion matrix defined below. We will first summarize overall performance and then analyze how performance depended on the precise shape of different randomly generated scenes.

We then reduce the data further by collapsing across the four directions of movement. Before doing that, we have to demonstrate that the observers’ patterns of error are isotropic. We then collapse the data for each observer into three measures, the frequency of a correct response ($0^\circ$ error), the frequency of a misclassification that is $90^\circ$ in error (e.g., a response “left” when up is correct), and the frequency of a misclassification that is $180^\circ$ in error (e.g., a response “down” when up is correct). The $180^\circ$ errors are of particular interest since they could be due to the ambiguity illustrated in Figure 2. If, for example, the rate of $180^\circ$ errors is low compared to the rate of correct responses, we will be able to conclude that observers are successful in combining shape and shading information to estimate the direction of motion of the collimated source.

Overall performance

Mean hit rate (between 0 and 1 inclusive) averaged over the four motion directions varied across observers from 0.42 to 0.84, with mean $\pm 1$ SEM equal to 0.65 $\pm 0.13$, which indicates that all observers performed above chance (0.25) in the experiment, as shown in Figure 6A.

Each observer’s 1,200 classifications were summarized in a $4 \times 4$ confusion matrix, where each row corresponded to a motion direction (left, right, up, down), and each column to a classification (“left,” “right,” “up,” “down”). The entry of each cell, $(d, R)$, was the total number of trials on which the observer classified veridical motion direction $d$ as motion $R$. Considering each veridical motion direction separately (300 trials total per row),
observers classified the axis (up–down or right–left) of motion correctly on 77 ± 11% (mean ± 1 SEM) of the trials, a proportion significantly higher than chance (50%) for all observers at all motion directions, except for S01, who was at chance to discriminate the motion axis throughout the experiment. Although counterintuitive, S01 was still above chance at discriminating motion direction, as all values on the confusion matrix diagonal (range = [113, 127]) were significantly greater than chance (75) at the \( \alpha = 0.05 \) level as assessed with binomial confidence intervals. As previously stated, all other observers were also significantly better than chance at classifying the motion direction.

Performance across the 300 scenes

Performance varied across the 300 scenes we generated. Each scene can be characterized by its anisotropy: some scenes supported horizontal light motion discrimination better than vertical or vice versa, which we term the scene’s response-based anisotropy. For scene \( i \), the response-based anisotropy, \( A^i_R \), is computed as the general increase in performance seen on trials with horizontal light paths, as opposed to vertical light paths:

\[
A^i_R = (N^i_{\text{right}} + N^i_{\text{left}}) - (N^i_{\text{up}} + N^i_{\text{down}}),
\]

where \( N^i_d \) is the total number of correct classifications across all subjects (maximum possible = 13) for scene \( i \) under light path \( d \). Many scenes had response-based anisotropies close to zero, but some were more extreme. See the histogram over all 300 scenes in Figure 6B. To explore the physical features of the scenes correlated with extreme response-based anisotropies, we plot the surface contours of the twenty scenes with the most extreme response-based anisotropies in Figure 7. As we might expect, the scenes for which horizontal light movements are more accurately judged tend to have roughly vertical ‘ridges’ and ‘valleys’ while those for which vertical light movements are more accurately judged tend to have elongated horizontal features.

To characterize this anisotropy physically, we computed the relative power in the horizontal frequencies versus the vertical frequencies at the first ten frequencies in units of cycles per image. To do so, we computed the Fourier Transform of the shape map of every scene separately. We developed a method more sensitive than the 2D Fourier Transform at characterizing these anisotropies, which were often evident as wood–grain like patterns of either horizontal or vertical power (Figure 7). We calculated the 1D Fourier Transform on every row of the \( z \) map image, stored its power spectrum, effectively throwing away phase information, and then summed over all rows for the horizontal power spectrum of the entire \( z \) map. We calculated the vertical power spectrum in the same fashion, working across columns instead. The physically based anisotropy of scene \( i \), \( A^i_f \), at each frequency \( f \) was then:

\[
A^i_f = p^{i, \text{horizontal}}_f - p^{i, \text{vertical}}_f,
\]

where \( p^d_f \) is the summed 1D Fourier power in the \( d \) dimension at frequency \( f \).

To test whether the response-based anisotropy of a scene could be predicted by the physical anisotropy, we performed the following regression:

\[
A^f_R = \sum_{f=1}^{10} a_f A^i_f + a_0.
\]

The regression was significant, \( F = 23.9, p < 0.001, R^2 = 0.45 \). A plot of the relative contribution of each frequency in cycles per image is shown in Figure 8. The analysis reveals that relative power differences at low spatial frequencies (1–3 cycles per image) are most predictive...
of anisotropies in classification performance. We will refer to this result later in the evaluation of our generative models.

**Sensitivity for different motion directions**

Sensitivity within each motion direction was also estimated for each observer separately by $d'$ from signal detection theory, a measure that is independent of observer bias (Green & Swets, 1966/1973). For a particular motion direction, we considered signal trials as those with light paths in the correct direction, and all remaining trials as non-signal trials. Therefore, we defined the hit rate for a particular motion direction, $p_H$, to be the probability of a correct motion classification, and the false alarm rate, $p_F$, to be the probability of classifying a trial as that particular motion direction when a different motion had occurred. If $\Phi^{-1}$ is the inverse of the cumulative unit normal distribution, then

$$
d' = \Phi^{-1}(p_H) - \Phi^{-1}(p_F). \tag{7}
$$

A zero value of $d'$ indicates chance performance, and $d'$ increases with increased discrimination performance. In
observers, as evidenced by sensitivity to each motion direction was above chance for all constructs 95% confidence intervals. The 5th and 95th percentiles were calculated and used to correspond to the correct direction. The pattern of misclassifications is likely as misclassifications in the opposite direction, due to the nature of the shading ambiguity for concave and convex surfaces (Figure 2).

our task, a \( d' = 1 \) corresponds to 69% correct, \( d' = 2 \) corresponds to 84% correct, and \( d' = 3 \) corresponds to 93% correct if a symmetric criterion is adopted. Ninety-five percent confidence intervals for each \( d' \) estimate were obtained by a non-parametric bootstrap method (Efron & Tibshirani, 1993): each observer’s performance in the corresponding condition was simulated 100,000 times and the 5th and 95th percentiles were calculated and used to construct 95% confidence intervals.

As expected on the basis of the confusion matrix analyses, sensitivity to each motion direction was above chance for all observers, as evidenced by \( d' \) 95% confidence intervals that did not include zero, and was moderately high on average with \( d' \pm 1 \ SEM \) equal to 1.6 ± 0.56 across observers.

The 13 observers’ motion direction sensitivities were close to isotropic across the four cardinal directions. Two observers showed significant anisotropies: S03 was more sensitive to upward changes (\( d' = 1.5 \)) than rightward (\( d' = 1.2 \)), and S10 was more sensitive to both upward (\( d' = 2.0 \)) and downward motions (\( d' = 2.1 \)) compared to leftward motion (\( d' = 1.6 \)). Significance was assessed with z-tests at the \( \alpha = 0.05/4 \) level, as we corrected for the four tests on each observer’s classification data.

**Misclassifications**

The analysis we now focus on regards misclassifications of motion direction. The pattern of misclassifications is important because it reveals the sources of error in the task: errors in shape or shading. As discussed above, there are two types of misclassifications possible, 90° errors, in which an adjacent classification category is selected, or 180° errors. Each observer’s pattern of 90° and 180° errors is a signature of that observer’s classification performance where, for example, a high proportion of 180° errors suggests depth map errors due to ambiguity.

Because sensitivity was relatively isotropic across the four directions of change, we combine across the four directions to report misclassification patterns. In general, observers committed higher rates of 90° errors than 180° errors (median ratio of 90°:180° errors = 2.0). These are shown in Figure 9 as black bars. Notice, however, that the proportion of correct response (“0°”) is higher for all observers than the proportion of 180° errors. On those trials where observers correctly judged the axis of change, they were between 2 and 36 times more likely to select the correct direction (median = 5 times more likely). Because concavities and convexities were equally as likely as in our scenes, the shading patterns alone were ambiguous with respect to lighting direction. This result confirms that our observers utilized shape information to estimate the light motion. If they had not, hit rates would be equally as likely as misclassifications in the opposite direction, due to the nature of the shading ambiguity for concave and convex surfaces (Figure 2).

**Model**

A well-supported claim in the literature is that human observers rely on luminance-based gradients to compute lighting direction: as a consequence of Lambert’s Law, the luminance gradient points to the light source if the surface is convex (Koenderink & Pont, 2003; Koenderink et al., 2003; Pentland, 1982). Pentland (1982) developed an algorithm based on the assumption of isotropically distributed surface gradients, which recovers illumination direction from single images using image intensity gradients. Using a small sample of photographs of natural convex objects, he found that estimates based on this algorithm predicted both the variability and errors of human estimates. Because of the ambiguity discussed in the Introduction section, this algorithm will fail if the surfaces in the scene portrayed in the image are not predominantly convex.

Koenderink et al. (2003, 2004) have also demonstrated that the pattern of errors human observers make when viewing monocular images and classifying the azimuth of the illuminant in the absence of cast shadows is consistent with a strategy of relying on the first-order directional derivatives of the luminance map of an image. As a consequence, when viewing images of surfaces containing both concave and convex regions from a monocular view, human observers err in estimating azimuth by 180° on about half of trials.

However, in binocularly viewed scenes, observers have access to information about shape, including concavity and convexity, in the form of shape estimates based on binocular disparity. The results of the experiment just presented demonstrate conclusively that they use this...
information to disambiguate their estimates of change in light direction. However, how exactly do they combine shape and shading information?

We present a new model that is an extension of Pentland’s and test its appropriateness as a model of human performance. Specifically, we use the model to predict overall performance and the patterns of human misjudgments summarized in Figure 9. We also use it to predict the anisotropy of the scenes as defined in the preceding section.

Our model works on two static inputs: the first is the shading intensity map, $I$, of the scene, and the second is the shape map of the scene. The model is illustrated schematically in Figure 10. The output of the model is a three-dimensional vector, $\hat{L}_{xyz} = (\hat{L}_x, \hat{L}_y, \hat{L}_z)$, pointing in the direction of the illuminant. The angle of the vector $\hat{L}_{xy} = (\hat{L}_x, \hat{L}_y)$ counterclockwise above positive $x$ modulus $360°$ is an estimate of the azimuth component of the illuminant:

$$\hat{\psi} = \tan^{-1}\left(\frac{\hat{L}_y}{\hat{L}_x}\right).$$

The angle of $\hat{L}_{xyz} = (\hat{L}_x, \hat{L}_y, \hat{L}_z)$ above the $xy$-plane is an estimate of the elevation of the illuminant:

$$\hat{\phi} = \tan^{-1}\left(\frac{\hat{L}_z}{\|\hat{L}_x, \hat{L}_y\|}\right),$$

where $\|v\|$ denotes the usual vector norm.

To estimate the change in lighting direction projected on the $xy$-plane, $\Delta L_{xy}$, from the model outputs, we simply

Figure 9. Patterns of responses. Observers’ responses are classified as correct ($0°$ error), $90°$ in error, or $180°$ in error. The bar plots summarize the relative frequency of each type of response.
subtract the initial estimate of direction (first frame) from the final estimate (last frame):
\[
\Delta L_{xy} = L_{xy}^f - L_{xy}^i. \tag{10}
\]
We categorize the resulting estimate \(\Delta L_{xy}\) as shown in the inset to Figure 10 to classify the change in lighting direction as one of the four cardinal directions: left, right, up, down.

### Shading map input

The shading processing component of the model is a direct implementation of Pentland’s (1982) model, in which illumination direction is estimated from the average of the shading gradients across the image, the maximum likelihood estimate when one assumes that the distribution of surface orientations in the scene is isotropic, an inaccurate characterization of the surface orientations in our scenes, which could be globally skewed toward either concavity or convexity. We summarize Pentland’s method here but note that we do not compute estimates of illuminant direction until after performing the shape correction described in the following section. After calculating the shading intensity gradient, \(\nabla I = (dI_x^r, dI_y^r)\), in the \(x\) and \(y\) directions at every point \((r, c)\) across an image, initial estimates of the \(x\) and \(y\) components, \((\hat{L}_x, \hat{L}_y)\), of the illuminant direction are computed as the mean gradient components under Pentland’s derivation:
\[
(\hat{L}_x, \hat{L}_y) = (\overline{dI_x}, \overline{dI_y}). \tag{11}
\]

Estimation of the \(z\) component of the illuminant direction requires two further calculations, which Pentland derives (but see also Chojnacki, Brooks, & Gibbons, 1994). We do not describe them here because our task required estimating only the relative magnitudes of the \(x\) and \(y\) components of the illuminant direction.

### Shape map input

Our observers were also provided with veridical depth information from binocular disparity. We have shown in previous psychophysical experiments that human observers utilize binocular disparity to correct the shading gradient direction when estimating the illuminant direction (Gerhard & Maloney, 2010). Thus, we included the shape map in our model, which was used to correct the local gradients in concave regions. Simply put, if a region is concave, the luminance gradient points in the opposite direction to the light source. A simple solution to this
problem is to flip the direction of the gradient in locally concave regions:
\[
\nabla l_{\text{concave}} = -\nabla l = (-dl_x, -dl_y).
\]

As an estimate of local shape, we computed Koenderink’s Shape Index, \( S \), (Koenderink, 1990, p. 320), which varies continuously between \(-1\) and \(+1\) and represents the shape (from convex elliptical to convex hyperbolic to flat to concave hyperbolic to concave elliptic) but not the degree of curvedness of the local region. It is derived from local estimates of principle curvatures and is negative for concave regions and positive for convex regions. By computing the Shape Index, \( S^{\epsilon,c} \), at every point \((r, c)\) in the image plane, we have a map varying from \(-1\) to \(+1\) that specifies the local shape of every visible point on the landscape. Taking the sign of this map, \( \text{sign}(S^{\epsilon,c}) \), simplifies the representation, allowing us to encode only concavity (\(-\)) and convexity (+).

Applying an element-wise multiplication of this concavity map, \( K^{\epsilon,c} = \text{sign}(S^{\epsilon,c}) \), with the map of local luminance gradients effectively flips all the gradients in the concave regions and leaves the convex regions undisturbed:
\[
\nabla l_{\text{corrected}}^{\epsilon,c} = K^{\epsilon,c} \circ \nabla l^{\epsilon,c} = K^{\epsilon,c} (dl_x^{\epsilon,c}, dl_y^{\epsilon,c}),
\]
where \( \circ \) denotes element-wise multiplication of matrices. After applying this correction, the estimate of lighting direction from Equation 11 was computed as follows:
\[
\langle L_x, L_y \rangle = \frac{\sum \epsilon,c K^{\epsilon,c} (dl_x^{\epsilon,c}, dl_y^{\epsilon,c})}{RC},
\]
where \( R \) is the total number of rows in the image and \( C \) is the total number of columns.

**Derivative filter**

There are several possible choices of filter shape and scale for computing the shading intensity gradients, \( (dl_x^{\epsilon,c}, dl_y^{\epsilon,c}) \). It is reasonable that information from the most informative scale, which depends solely on the spatial variation in the stimulus, would be the scale of interest. We consider the questions of “which derivative filter?” and “which scale?” empirically less important for this analysis, which is the first of its kind, and choose instead to focus on the predictive power of the basic form of the model. We have therefore chosen a scale and a derivative filter demonstrated to have high accuracy, the 5-tap filter developed by Farid and Simoncelli (2004), which is shown in the gradient box of Figure 10. The size of the filter on our stimuli corresponds to 0.125 DVA on a side.

This choice of filter also solves a minor problem in computation for us. The albedo variation in the small triangular facets used in rendering is effectively eliminated by the 5-tap filter, and we can directly use the luminance map in place of a shading map (the luminance variation present if the scene were homogeneous in albedo). If we wished to apply the model to scenes with large patches differing in albedo, we would need to first remove all albedo variation to produce a true shading map.

**Effect of the diffuse light source**

The effect of the diffuse source, which was one-quarter the intensity of the collimated light source (see the Methods section), is to scale the luminance gradient. It does not affect the direction of the gradient.

**A note on circular data**

To evaluate this model, we routinely work with angular estimates of lighting azimuth computed using Equation 8 and with veridical azimuth values also encoded as angles. In the following analyses, all statistics on circular data were computed using the circStat2009 MATLAB toolbox (Berens, 2009).

**Noise-free model applied to static scenes**

We remind the reader that our observers did not report estimates of the illumination azimuth on static scenes but instead were required to report only the direction of change in lighting.

The model cannot recover illumination azimuth perfectly on every static image. We estimated lighting azimuth on all 38,400 left and right eye images from the experiment and compared with the known azimuth values. We found that the estimator is practically unbiased (mean error = \(+0.4^\circ\)) and has a circular standard deviation of 28.0°. The interquartile interval of the deviations of the estimates from the true azimuth values was \(\pm 12.1^\circ\), which is remarkably similar to the interquartile deviation ranges of human observers in estimating illuminant azimuth from Gaussian surfaces, which are in the range \([\pm 7.1^\circ, \pm 13.7^\circ]\) apart from 180° errors (Koenderink et al., 2004). We also analyzed the left and right eye images separately and found no practical differences between the associated estimates.

One should note that the static noise is not necessarily a determinant of the noise in the illuminant change direction estimates because the motion estimates are based on differences between frames, where common noise due to the static aspects of the scenes falls out of the computation. Noise in static scenes is presumably largely due to the surface shape, albedo texture, tessellation, and the finite size of the 5-tap operator.
Noise-free model applied to predict change in lighting

We characterize the noise-free model’s classification performance of lighting change under two cases: (1) without shape information, and (2) with the shape correction introduced. The first case corresponds to Pentland’s (1982) model. The second case corresponds to our experiment, in which binocular disparity disambiguated shape.

Without shape information, the model could classify the axis of change (up–down vs. left–right) correctly on 70% of the trials (843/1200). Within those trials, 52% (439/843) were classified with the correct direction, and 48% classified in the opposite direction (404/843). In other words, although the model was well above chance to classify the axis of motion, it was at chance in classifying the direction, as expected. The remaining 30% of errors were classifications into the adjacent category, or 90°-errors. Figure 11A contains rose plots of the noise-free model’s change direction estimates for all 1,200 trials separately for each true motion direction.

These results illustrate the pattern of errors expected if one were computing illuminant direction without taking the surface shape into account. In other words, the change in each trial’s mean shading gradient from frame 1 to frame 16 is the sole determinant of the motion direction estimate in this analysis. Because our scenes were made up of convex elliptic, concave elliptic, and hyperbolic (saddle) regions, the accuracy of the mean luminance gradient depends on the relative proportion of such regions; it has been noted that saddle regions can generate luminance gradients orthogonal to the true lighting direction (Koenderink & Pont, 2003), and we have already seen that concave elliptic regions generate luminance gradients opposite in direction to the true lighting direction. Together these effects generate the pattern of errors shown in Figure 11A.

This pattern of results qualitatively replicates human errors from previous work, in which rendered surfaces were presented in a single image viewed binocularly (Koenderink et al., 2003, 2004). We have also replicated these results with human data (Figure A1), in a control experiment, in which we presented monocular views of the same stimuli (described in the Discussion section and Appendix A).

When the correct depth map is provided to the model, and our concavity correction is applied, the estimates of motion direction lead to perfect classification of every trial. Rose plots of the change direction results are shown in Figure 11B. The standard deviation of the estimates for each motion direction is on average 7.8°, and the estimates are virtually unbiased.

**Evaluation of the model**

The goal of this work is to model human lighting motion classification performance based on the first-order directional derivatives of the luminance map, subject to
correction for concave regions. Estimating illuminant direction from first-order directional derivatives of the luminance map is not a novel idea (Knill, 1990; Lee & Rosenfeld, 1985; Pentland, 1982) but correcting the derivatives for the measured local 3D shape information is. Because we have shown in previous psychophysical experiments (Gerhard & Maloney, 2010) and in the experiment reported here that human observers correct for concave regions when judging lighting direction in stereoscopic scenes, we included a concavity correction in the model.

The noise-free model can classify every trial in our experiment correctly, indicating that each trial contains enough information to perform the motion classification. In fact, each trial is a wealth of redundant information from a computational perspective, as the model, which can detect even a very small change of mean gradient, can still classify all 1,200 trials correctly when limited to a central patch subtending 0.25 DVA on a side. Human performance, on the other hand, had an effective ceiling of 84% correct.

In comparing our human results with the model’s classification performance without shape information, we find that the human data are remarkably better, indicating once again that the observers incorporated the shape information in their judgments. Our observers were above chance to detect change direction correctly, as evidenced by the much larger hit rates as compared to 180° errors (see Figure 9).

Unlike our noise-free model, the human visual system is subject to multiple sources of internal noise, optical, neural, decision-based, and memory-based, all of which could potentially degrade the signals present in our trials. In this evaluation, we focus on the effects of internal noise in representing the shading and depth maps of our scenes, with the goal of predicting not only the correct classifications of our observers but also the pattern of errors they made (Figure 9).

To test for effects of noise, we selectively add noise to one input of the model, e.g., the shape map or the shading map while holding the other input correct. By simulating ideal observers at various levels of noise in one input map, we can then compute maximum likelihood estimates of each observer’s respective noise level, and finally compare the observer’s overall pattern of performance (0°, and 90°, 180° error rates) with his matched ideal observer’s pattern. We will consider an ideal observer a good model of the observer only if the overall pattern matches, and the maximum likelihood estimate of internal noise is plausible in magnitude.

**Luminance degradation model**

In this analysis, we hold the shape map correct and perturb the luminance information to varying degrees by the addition of 2D pink (1/f) noise. The spectra of our stimuli were roughly 1/f, so we expected pink noise to be effective in disturbing the luminance patterns. We further speculate that the shape of pink noise may closely resemble internal noise if spatial frequency filter outputs are normalized and then followed with additive noise.

**Pink noise samples**

We simulated each sample of pink noise by starting with an image of Gaussian noise sampled from MATLAB’s randn function. We then filtered the image in the frequency domain by a rotationally symmetric two-dimensional 1/f filter and took the real part of the inverse 2D Fourier Transform, which was then normalized to contain pixels with values on the interval [−1, 1]. To specify a particular level of noise in the same units as the luminance maps of the trials (cd/m²), we multiplied the pixels of the noise image by a gain g specified in cd/m². Under sufficiently high gains, adding the noise sample to a trial’s luminance map could result in negative pixel values. When this occurred, negative pixel values were clipped to zero.

**Simulations**

To simulate luminance noise on a trial \( t \), we added independent samples of pink noise to the first and last frames of each trial. We then processed both frames with the model and categorized the resulting change in lighting direction. This was repeated \( N \) times for each trial \( t \). The resulting \( N \) simulated directions were stored in \( d_n \), where the \( n \)-th entry indicates the lighting change direction \( d \in \{1: \text{“up,” 2: “down,” 3: “right,” 4: “left”}\} \) of the \( n \)-th simulated trial. The proportion of responses equal to \( d \) were stored as \( P_{g,d}(d) \), the Monte Carlo probability of classifying trial \( t \) as direction \( d \) under noise level \( g \):

\[
P_{g,d}(d) = \frac{\sum_{n=1}^{N} \left[ \hat{d}_n = d \right]}{N},
\]

where \( \hat{d}_n = d \) has value 1 if left- and right-hand terms

![Figure 12](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933478/)
match, otherwise 0. The likelihood matrix \( \mathbf{L} \) was computed by taking the natural log of \( \mathbf{P} \):

\[
\mathbf{L}_{g,t} = \ln(\mathbf{P}_{g,t}).
\]

This procedure was repeated at multiple noise gain levels.

**Maximum likelihood estimates of noise**

The data of each observer are contained in a vector of 1,200 classifications, \( c_t \), where the \( t \)-th entry is the direction \( d \in [1: \text{“up,” 2: “down,” 3: “right,” 4: “left”}] \), which the observer selected on trial \( t \). The likelihood of the data under gain \( g \), \( \hat{L}_g(c) \), is the sum of each classification’s likelihood under gain \( g \):

\[
\hat{L}_g = \sum_{t=1}^{1200} \mathbf{L}_{g,t}(c_t).
\]

The MLE of noise gain is then

\[
\hat{g} = \arg\max_g (\hat{L}_g).
\]

**Results**

We first wish to emphasize that the MLEs for the 13 observers corresponded to very large luminance noise levels. Indeed, high levels of noise were required before the model deviated at all from perfect classification performance. Under simulated pink noise with gain equal to 100% of the mean luminance, the model classified 99.8% of the trials correctly. In order to reach performance levels comparable to that of the observers, gains of at least 388% of the mean luminance were required. At the highest end, luminance noise was estimated to have a gain of 713% of the mean luminance. At these levels of luminance noise, between 14% and 28% of the pixels were clipped to zero during simulations. The MLE for each observer is printed above each one’s plot in Figure 12A.

Moreover, we could not reproduce the patterns of errors in observers’ data with only luminance noise. In Figure 12A, we replot each observer’s signature pattern of responses from Figure 9: hit rate, 90° error rate, and 180° error rate as the black bars with 95% binomial confidence intervals. For each observer, we also plot the matched ideal observer’s signature pattern, in dashed gray. The matched ideal observer’s signature pattern was computed from \( \mathbf{P}_g \), where the gain level \( g \) was the observer’s MLE.

On average, the maximum likelihood fits underpredicted hit rates by 0.085, a significant underestimation, \( t(12) = -5.1, p < 0.001 \). Predicted 90° error rates were highly inflated for all but one observer and were on average 0.14 greater than achieved 90° error rates, a significant effect, \( t(12) = 5.7, p < 0.001 \). Predicted 180° error rates, on the other hand, were underestimated by the maximum likelihood fits by 0.054 on average, a significant effect, \( t(12) = -5.1, p < 0.001 \).

The large and patterned disparities between observers’ overall hit rates and respective matched ideal observer hit rates do not result from insufficient gain dimension sampling. The gain values we simulated (100%–712.5%) were sampled in steps of 12.5%, which corresponded on average to 0.008 increases in overall hit rate (between 0 and 1 inclusive). In other words, the simulated ideal observers varied in overall hit rate in steps of less than 0.01 on average, and at most the increase between two simulated observers was 0.017, demonstrating that all simulated ideal observers spanned the range of our observers’ performance with high resolution. The disparities we observed resulted from the maximum likelihood fitting procedure on each observer’s vector of classifications. Not one observer’s signature of performance was well fit by his matched ideal observer, operating under luminance noise equal to the observer’s MLE of noise gain.

**Discussion**

We reject the luminance noise model of our observers’ performance, for multiple reasons. First, we found that very large levels of noise were necessary to simulate observers’ classifications trial by trial. Even at the lowest end, MLEs of noise gain were near 400% of the mean luminance. This indicates that the luminance maps must have been degraded substantially before the model would select the same motion direction on a trial that an observer did. Second, even if the MLEs could be considered as reasonable levels of equivalent internal noise on the luminance map representation, a model with noisy luminance map representations cannot predict the signature patterns of performance. Specifically, noisy luminance maps predict a pattern of errors where motion estimates are much more likely to fall into adjacent categories, 90° errors, than was observed, and they predict fewer 180° errors than observed. Third, we have repeated this luminance noise analysis using white (Gaussian) noise and found similar results both in terms of noise levels required (noise standard deviation = 100%–295% of the mean luminance) and the signature pattern of performance, which also overestimated 90° errors and underestimated 180° errors, in addition to underestimating the hit rates on average.

There are several reasons why this model is highly tolerant of luminance noise. First, the model performs two-dimensional directional derivatives over every pixel in the image (excluding boundaries) with a filter of size of \( 5 \times 5 \) pixels. To the extent that the noise is spatially correlated at a local scale, the derivative operation will be highly robust to it, and \( 1/f \) noise is spatially correlated. Second, the model averages derivative responses across the image separately within each dimension, so the final estimate of static light direction is also robust to noise in
the individual derivative responses. Finally, the motion direction estimate is computed as a difference in the model’s estimate between two frames.

In conclusion, we cannot explain errors in rapid judgments of illumination direction as primarily due to degradation of the luminance map signal.

Shape map degradation model

In this analysis, we do not perturb the shading maps but deform the shape map smoothly and randomly to varying degrees by perturbing the underlying Gaussian height parameters. Each scene was an aggregate of 30 Gaussians, each with a random height. We simulated noisy scenes by generating new aggregate scenes after adding noise to the true Gaussian heights.

Simulations

Each scene was generated as an aggregate of 30 isotropic Gaussians, whose heights were specified by \( h_i \) (see Equation 1). To simulate depth noise at level \( \sigma \) cm, we sampled 30 height noises, \( n_i \), from a normal distribution:

\[
n_i \sim N(0, \sigma),
\]

and then generated a new scene with shape map specified by Equation 20 where heights, \( h_i \), have been perturbed by, \( n_i \):

\[
z(v) = \sum_{i=1}^{30} (h_i + n_i) \frac{1}{(2\pi|\sum|^1/2)} e^{-\frac{1}{2}(v-\mu)^T \sum^{-1}(v-\mu)}.
\]

As before, \( v = (x, y) \) denotes any point in the display area. For each scene \( s \), this process was repeated \( N \) times to generate \( N \) smoothly deformed versions of scene \( s \) resulting from Gaussian height noise normally distributed with standard deviation \( \sigma \) cm.

Each scene was rendered under four light paths, so the simulated shape maps of a scene were applied to each of the four trials linked to that scene. To simulate depth noise on a trial \( t \), we processed the first and last frames of each simulated trial and categorized the change in lighting direction. The resulting \( N \) simulated directions were stored in \( \tilde{d}_n \), where the \( n \)th entry indicates the direction \( d \in \{1: \text{“up,”} 2: \text{“down,”} 3: \text{“right,”} 4: \text{“left”}\} \) of the \( n \)th simulated trial. The proportion of responses equal to \( d \) were stored in \( \mathbf{P}_{\sigma,t}(d) \), the Monte Carlo probability of classifying trial \( t \) as direction \( d \) under noise level \( \sigma \):

\[
\mathbf{P}_{\sigma,t}(d) = \frac{\sum_{n=1}^{N} \tilde{d}_n = d}{N},
\]

where \( \tilde{d}_n = d \) has value 1 if left- and right-hand terms match, otherwise 0. The likelihood matrix \( \mathbf{L} \) was computed simply by taking the natural log of \( \mathbf{P}_{\sigma,t} \):

\[
\mathbf{L}_{\sigma,t} = \ln(\mathbf{P}_{\sigma,t}).
\]

This procedure was repeated at multiple depth noise levels.

When the depth map is perturbed by sufficient noise, local shape indices are also perturbed, and these changes in local shape index are responsible for errors in classifying the direction in which the light moved. That is, if the concavity/convexity labels do not change, then the model would never make an error in the task. For any given surface, some local shape indices are more likely to change with added depth noise, but the link between surface structure, vulnerability, and error in judging light direction remains to be analyzed.

Maximum likelihood estimates of noise

The data of each observer are contained in a vector of 1,200 classifications, \( c_t \), where the \( n \)th entry is the direction \( d \in \{1: \text{“up,”} 2: \text{“down,”} 3: \text{“right,”} 4: \text{“left”}\} \), which the observer selected on trial \( t \). The likelihood of the data under noise level \( \sigma \), \( \hat{L}_\sigma(c) \), is the sum of each classification’s likelihood under level \( \sigma \):

\[
\hat{L}_\sigma = \sum_{t=1}^{1200} \mathbf{L}_{\sigma,t}(c_t).
\]

The MLE of noise is then

\[
\hat{\sigma} = \arg\max_{\sigma}(\hat{L}_\sigma).
\]

Results

MLEs for the 13 observers corresponded to a small range of depth noise levels, \( \hat{\sigma} \in [0.80, 0.90] \) cm. The MLE for each observer is printed above each one’s plot in Figure 12B, in which we show each observer’s signature pattern of responses: hit rate (0° error), 90° error rate, and 180° error rate as the black bars with 95% binomial confidence intervals. The calculation of these rates was explained in the Results section above. For each observer, we also plot the matched ideal observer’s signature pattern, in dashed gray. The matched ideal observer’s signature pattern was computed from \( \mathbf{P}_\sigma \), where the noise level was the observer’s MLE. On average, the maximum likelihood fits predicted unbiased hit rates (-0.01) and 90° error rates (-0.01), both not significant. Predicted 180° error rates were slightly biased by +0.02, an effect that was significant, \( t(12) = 2.18, p < 0.05 \), yet of a very small and therefore negligible magnitude.
Discussion

We conclude that uncertainty in shape maps can account for the patterns of errors of human observers. The estimated error sizes were reasonable and fell in a narrow range across all observers, and signatures of performance matched.

Our model of surface depth degradation assumes that observers have internal estimates of the underlying bump heights that are noisy. While there is no empirical support in the literature on this particular conjecture, we find it reasonable because our scenes were smooth surfaces and would be easily interpolatable even if large patches were missing (based on previous unpublished results in our laboratory), so we conjecture that errors in estimating depth across the surface should be spatially correlated. The good agreement of our model with human data suggests that our assumption is not unreasonable.

To test the validity of our model’s assumption, we tried a second version of depth map degradation, which was equivalent to adding noise matched in the frequency domain for our shape maps while being agnostic about bump locations in space. We found that it required far higher noise levels (>3 cm) before it could approximate observers’ classification performance and did not lead to the patterns of errors found in human data. These results indicate that processing the pattern of bumps in the shape map and combining it with the luminance information is the major source of error in judgments of illumination direction change.

In Appendix A, we report a control experiment demonstrating that when observers attempted to classify lighting direction without binocular disparity information, they made 180° errors equally as often as correct classifications (Figure A1). This outcome confirms that it is precisely binocular disparity information that observers used to judge shape.

In summary, the shape model fits matched the pattern of errors across observers better than the shading model did. The shape model fits across observers were unbiased for 0° and 90° errors and on average overstated the 180° error rate but only by +0.02 whereas the shading model largely underestimated the hit rates (−0.09), overestimated the 90° errors (+0.14), and underestimated the 180° error rate (−0.05).

Behavioral results

We reported an experiment in which observers estimated the direction of motion of an out-of-view collimated light source illuminating a three-dimensional scene viewed stereoscopically. The scene consisted of a three-dimensional textured surface composed of superimposed Gaussian bumps randomly varied in location and height. The task was a four-alternative forced-choice judgment, in which observers selected a direction of lighting change on each of 1,200 novel combinations of scene and light path.

The first point we wish to make is that the results of this experiment provide further evidence that observers have access to information about light sources in a scene that are not directly visible but rather inferred from the pattern of luminances across the scene. Our task precludes certain strategies available to observers in matching experiments, particularly when observers are matching shading appearance between objects, which has been the predominant task in the literature to assess illuminant direction estimation (Khang et al., 2006; Koenderink et al., 2003, 2004; Pont & Koenderink, 2007). In a matching task, particularly in matching two spherical objects, observers can match the direction to the illuminant between two scenes without having any estimate of the illumination. This could, for example, be accomplished by minimizing differences in luminance over regions of similar surface orientations.

Furthermore, our observers never received feedback. They were simply told to report the change in lighting direction, and the data indicate that all 12 naive observers, who spanned a large range of performances, were able to judge the change in direction above chance. We also strictly controlled the randomization of the light paths so that observers could not guess the correct response based on any one static frame of the experiment but were instead forced to update and remember their estimates over time.

In addition to being one of the first experiments to demonstrate a remarkable ability to infer and track small (10°) changes in lighting direction (see also Gerhard & Maloney, 2010), our results also demonstrate that the tendency to make 180° errors in estimating lighting direction is much reduced when scenes are viewed stereoscopically.

Model evaluation results

To test the appropriateness of luminance gradient models for human estimation of illuminant direction (Karlsson, Pont, & Koenderink, 2008, 2009; Knill, 1990; Koenderink & Pont, 2003; Lee & Rosenfeld, 1985; Pentland, 1982) in three-dimensional scenes that include concave regions, we implemented Pentland’s (1982) model and introduced a correction component, which addressed the gradient ambiguity that arises when surfaces can be concave or convex. The correction is contingent on the depth map being computed via the stereo disparity information. Our model can therefore be described as a Contingent Ideal Observer model of human performance. The concavity correction it performs is to process the depth map for local shape estimates indicating concavity and convexity. By flipping the gradient in concave regions...
before applying Pentland’s algorithm, the model can then recover the true lighting direction.

The static estimates of our Contingent Ideal Observer were as variable as human estimates of static lighting direction. In addition, without the concavity correction, the model committed 180° errors about as frequently as human observers would under monocular views of the scenes (see Appendix A). In our experiment, the stimuli were stereoscopic, so they contained a valid cue to depth, binocular disparity. Our results confirmed that observers utilize binocular disparity to incorporate shape information in their judgments although not perfectly.

We generated simulated Ideal Observer data by selectively adding noise to the inputs. In this manner, we compared the pattern of classification performance that resulted from shading noise versus shape noise. The results of the model evaluation indicated that shape degradation was a far more likely explanation of the observer’s pattern of results than shading degradation. Not only were the required levels of noise more reasonable under the shape noise Ideal Observer model, but the patterns of hit rates and 90° errors matched the subjects’ data, and 180° errors were highly consistent with the data. Injecting luminance noise into the shading maps could not replicate these error patterns and led instead to gross overestimates of 90° error rates and underestimates of both hit rates and 180° errors.

We hypothesize that there could be several reasons why our model evaluation resulted in a conclusion consistent with shape processing errors. First, solving the stereo depth problem is difficult and might not be well accomplished in a brief presentation such as ours where observers are focused on the dynamic luminance pattern. For example, the calculation we used to perform our concavity correction from the depth map was Koenderink’s (1990) Shape Index, which relies on a series of differential geometry computations involving second derivatives of the depth map. These computations might not be well accomplished by the human visual system under the constraints of our experiment. Second, we cannot logically separate errors in shape from errors in combining the shape and shading information. It is possible that the errors resulted from imperfect spatial binding between the shading map and the local shape map.

We also correlated performance anisotropy across the individual scenes with the anisotropy of the matched ideal observers’ classifications for both models. The matched shape degradation model observer explained 27% of the anisotropy variance, and the shading degradation model observer explained 28% of the anisotropy variance. Both R² were significantly different from 0, but the amount of variance accounted for is evidently modest. Further work is needed to explore the relationship between local shape and anisotropies in light field estimation.

The stimulus we presented to the human observers was a sixteen-frame stereo movie. However, the algorithm we developed to estimate the direction of movement of the light source made use of only the first and last frames. Even so, the model, given only these two frames, could outperform all human observers and, with noise added to the depth map, had a similar pattern of error to that of human observers. These results suggest that there is little useful information in the remaining frames, but, while that conclusion is valid for the particular conditions of our experiment, it is not true in general.

In our experiment, light sources moved only 10° along linear trajectories on the image plane (Figure 3, inset), and consequently, the effect of light movement on the two images of each stereo pair is nearly linear, with the intermediate frames well approximated by convex mixtures of the first and last frames. However, consider an experiment similar to ours but with light sources that can travel along curved trajectories where the observer’s task is to make judgments not simply about direction of movement but also the curvature of the trajectory of the light source.

The algorithm we developed, given access to all sixteen frames, is readily adapted to make such judgments. It could track the direction to the light source, frame by frame, as the light source moved along a curved trajectory. Based on these estimates it could return estimates of curvature. No algorithm given only the first and frames could do so: any light trajectories, curved or straight, that began and ended in the same location would be indistinguishable. Therefore, the intermediate frames carry essential information about light trajectory curvature and other properties of the trajectory.

However, can people make use of this information? Based on the results reported here, we know that observers can accurately classify direction of movement of light sources. Further research is needed to determine whether human observers can make accurate judgments about not just the direction in which light sources move but also the curvature and other properties of the trajectories they follow.

Appendix A

Monocular control experiment

Experiment

We repeated the experiment without the binocular disparity cue. We reran the same 1,280 trials (practice and test), but observers saw only the right or left eye image of each trial instead of both. In the main experiment of the paper, the trials were generated from 300 scenes rendered under 4 light paths each. In the control experiment, we presented the right eye image to both eyes for scenes 1–150 and the left eye image to both eyes for scenes 151–300. The trial order was randomized for each
subject. The task, timing, and apparatus were otherwise the same as described in the Experiment section. No feedback was given.

**Observers**

Five observers in the previous experiment participated in the control experiment and were reimbursed for their time at $10/h. We selected the five observers with the highest proportions of correct classification in the previous experiment (excluding the author HEG).

**Results and discussion**

Observers classified the axis of change correctly on 81 ± 2% of the trials. In this respect, observers performed better than the shape free model run on the same stimuli, which classified the axis correctly 70% of the time. This slight yet reliable improvement over the model in detecting the motion axis, and hence making fewer 90° errors, indicates that the human visual system is more sensitive to the luminance changes of our stimuli than our particular implementation of Pentland’s (1982) model. One possibility is that the human visual system’s effective derivative operation is an improvement over the 5-tap filter we applied (Farid & Simoncelli, 2004). A second possibility is that observers utilized the residual depth cues, such as texture compression and boundary contours, to estimate local shape correctly up to the sign, thus making concave/convex confusions but not saddle/elliptic confusions, which could constrain the probabilities of each motion direction to favor the true direction up to 180° if observers focused on the elliptic regions. The model did not have access to these residual cues and computed the mean across the entire scene, weighing saddle regions, which could contain gradients 90° away from veridical illuminant azimuth, equally as the more reliable elliptic regions.

While the observers were all significantly above chance in discriminating the axis of change, all were at or near chance in discriminating the direction of lighting change, which indicates that the residual depth cues were not enough to disambiguate concave regions from convex. Out of the trials on which the axis was classified directly, observers discriminated the direction correctly 51 ± 3% of the time, where chance performance is 50%. Each observer’s pattern of errors is plotted in Figure A1 along with their original results from the full 3D experiment for comparison in dashed gray.

The pattern of results using monocular stimuli replicates previous work demonstrating that observers randomly commit 180° errors in judging lighting direction when there is no shape information to disambiguate concavity and convexity (Koenderink et al., 2003, 2004).

**Acknowledgments**

We thank David Brainard, Eero Simoncelli, and Michael S. Landy for suggestions on an earlier draft. We thank James Elder, Felix Wichmann, Fabian Sinz, and
Matthias Bethge for helpful discussions. This work was supported by NIH/NEI EY08266.

Commercial relationships: none.
Corresponding author: Holly E. Gerhard.
Email: hgerhard@gmail.com.
Address: Spemannstraße 41, Tuebingen, 72076, Germany.

Footnotes

1A collimated light source can be thought of as a point source of light very distant from the scene (e.g., a star). The direction to the collimated source is the same at every point within the scene.

2We provide a detailed description in the Methods section for reproducing the stimuli. Interested readers may also contact the corresponding author for MATLAB code to reproduce the stereo stimuli on their own calibrated displays.

3We denote intervals using both square brackets and parentheses. The parenthesis is used when the endpoint of the interval is not in the interval, i.e., (a, b) does not include a or b but all the numbers in between. The square bracket is used when the endpoint is included, i.e., [a, b] includes a, b, and all the intervening numbers while (a, b] includes b but not a.

4MATLAB code for generating the filter is available from the authors at http://www.cs.dartmouth.edu/farid/research/derivative.html.

References


