Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes

Brian Potetz and Tai Sing Lee

Department of Computer Science, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Received October 3, 2002; revised manuscript received February 10, 2003; accepted February 10, 2003

In spite of the recent surge in the popularity of statistical approaches to vision, the joint statistics of coregistered range and light-intensity images have gone relatively unexplored. We investigate statistical correlations between images and the surface shapes that produced them. We determine which linear properties of range images can be best predicted from simple computations on intensity information, and we determine those properties of intensity images that best predict range information. We find that significant (up to $\rho = 0.45$) and potentially exploitable correlations exist between linear properties of range and intensity images, and we explore the structure of these correlations. © 2003 Optical Society of America

OCIS codes: 150.5670, 330.3790.

1. INTRODUCTION

The study of shape from shading has been a major branch of vision since the 1970s and has its roots even earlier, in the photometric investigations of the lunar surface performed in the 1920s.[1] Since this time, the standard approach to the problem has involved modeling the behavior of light as it travels through space and interacts with surfaces and then attempting to invert the image formation processes. Unfortunately, inverting this process is highly underconstrained, and various assumptions about image formation models and their parameters, such as Lambertian surface reflectance, uniform albedo, and shadow-free, single-point-source illumination, have to be made for this approach to work. However, these assumptions do not always hold in natural images, and this often leads to poor generalization for these algorithms.

There have been some notable exceptions to this trend. Shape-from-shading algorithms that make use of a direct association between luminance images and the three-dimensional (3D) models used to generate them were presented by Freeman *et al.*[2] and also by Lehky and Sejnowski.[3] In each of these cases the image statistics used were derived from computer-generated images. These images make the same assumptions made by traditional shape-from-shading algorithms. These approaches suggest that a deeper understanding of the joint statistics of natural images and their associated 3D structures might be important for the successful development of a statistical approach for 3D inference.

There has been considerable recent interest in the statistics of natural images, particularly in the context of image coding[4–7] and scale invariance[8,9] as well as in the joint statistics of neighboring pixels or wavelet responses.[10,11] However, there has been relatively little investigation into joint statistics of range and luminance images. We know of only one study that explores these joint statistics. It has been observed that human perception of the length of a line segment on a blank background depends on the orientation of the line. In a recent study, Howe and Purves[12] examine range images to find that this bias closely matches the 3D length of these line segments when projected into range images. The authors go on to restrict their investigation to only those intervals that correspond to edges in the luminance image and show that the finding still holds. This result shows how investigation into the statistics of range and coregistered luminance images may help to explain how depth is computed in the brain.

There may be many factors that affect the statistics of natural images that cannot be inferred from simple physical models. For example, the statistical relationship between images and their surfaces will be affected by the natural statistics of illumination direction, a factor that is known to heavily influence human performance on shape-from-shading problems.[13] Other factors that influence this relationship may include the statistics of object size and shape in natural scenes as well as the natural statistics of the surface properties of those objects. This opens up the possibility that there may be simple, exploitable statistical relationships between real images and surface shapes that have gone overlooked. Discovering these relationships might further the development of vision algorithms that utilize shape-from-shading information. It may also provide insight into how the human visual system is so adept at solving these problems.[14]

For this study we constructed a database of coregistered high-resolution two-dimensional (2D) color images and 3D range images using the most recent scanner technology. We then conducted a first attempt to explore the statistical correlations between images in these two domains. We searched for simple, local, linear correlations using linear regression, ridge regression, and canonical correlation. These techniques allowed us to extract some simple but interesting statistical trends between intensity and range images. We hope this and future exploration of these data will provide a better understanding of the statistical distribution and correlations between 2D images and 3D structures that will provide a solid foundation for an image-based statistical approach for 3D inference.

2. METHODS

To investigate these statistical relationships, we first collected a database of coregistered intensity and range images (where by coregistered we mean that corresponding pixels of the two images correspond to the same point in space). We selected the Riegl LMS-Z360, the most sophisticated available long-range scanner, to construct our image database. The Z360 collects coregistered range and color data by using an integrated color photosensor and a time-of-flight laser scanner with a rotating mirror. The scanner has a maximum range of 200 m and a depth accuracy of 12 mm. However, for each scene in our database, multiple scans were averaged to obtain an accuracy of 6 mm. For the purposes of this study, red, green, and blue color values were combined into one gray-scale light-intensity value. The measurement of each pixel is known to be independent. This means that no global image processing or automatic gain control is applied to the images. All range measurements are reported in meters. All scanning is performed in spherical coordinates.

Using this scanner, we collected scans of a wide variety of outdoor scenes. Example range and intensity image pairs appear in Fig. 1. All scenes in our database are outdoor shots, taken under sunny conditions during the week of June 17th, 2002, in western Pennsylvania. The camera was kept level with the horizon and was positioned either on the ground or on a tripod, roughly 1 m off the ground. Over 100 scenes were taken at various resolutions. However, for this paper we selected a 50-scene subset of our database with spatial resolution of 22.5 ± 2.5 pixels per degree. This removes from our dataset images with very high or low spatial resolutions. Extremely dark images were also discarded. The contents of our images include scans of trees and wooded areas, rocky areas, building exteriors, and sculptures. Twenty-one images were of urban scenes and twenty-nine were of rural scenes. Each image required minutes to scan, hence only stable and stationary scenes were taken. The average size of our images was 1000×604 pixels, for a total of 30,177,930 pixels.

Regions represented by monochrome white noise in Fig. 1 are areas where no range information is available. The surface could be out of range, such as sky, or too reflective to be measured by the scanner, such as water. All image patches that contain invalid pixels were excluded from our calculations.

As is common in the analysis of the natural statistics of images,[6,15] we work with the logarithm of the light-intensity values rather than intensity itself. One advantage of this is that image contrast, rather than being a

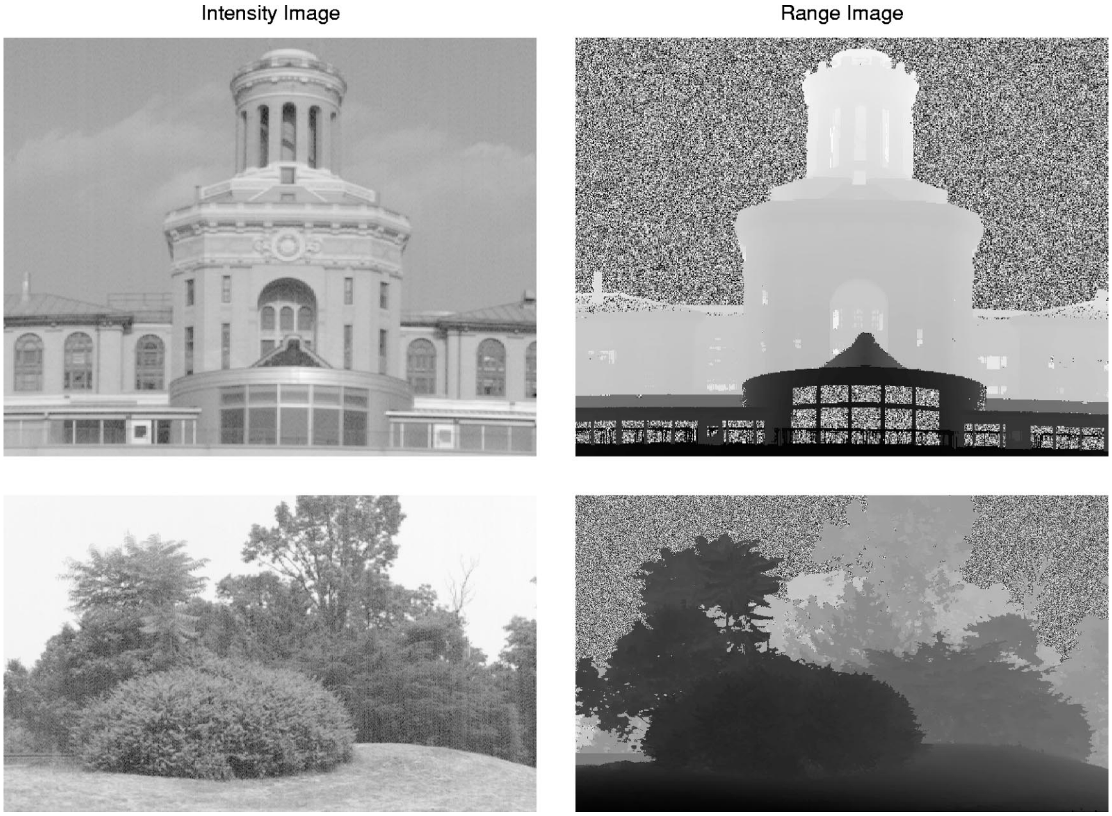

Fig. 1. Two sample image pairs from our database: one from the urban subset, one from the rural subset. The urban image is of Hammerschlag Hall, Carnegie Mellon University, Pittsburgh. The rural image is of trees and bushes in Schenley Park, Pittsburgh. Out-of-range regions are depicted using monochrome white noise.

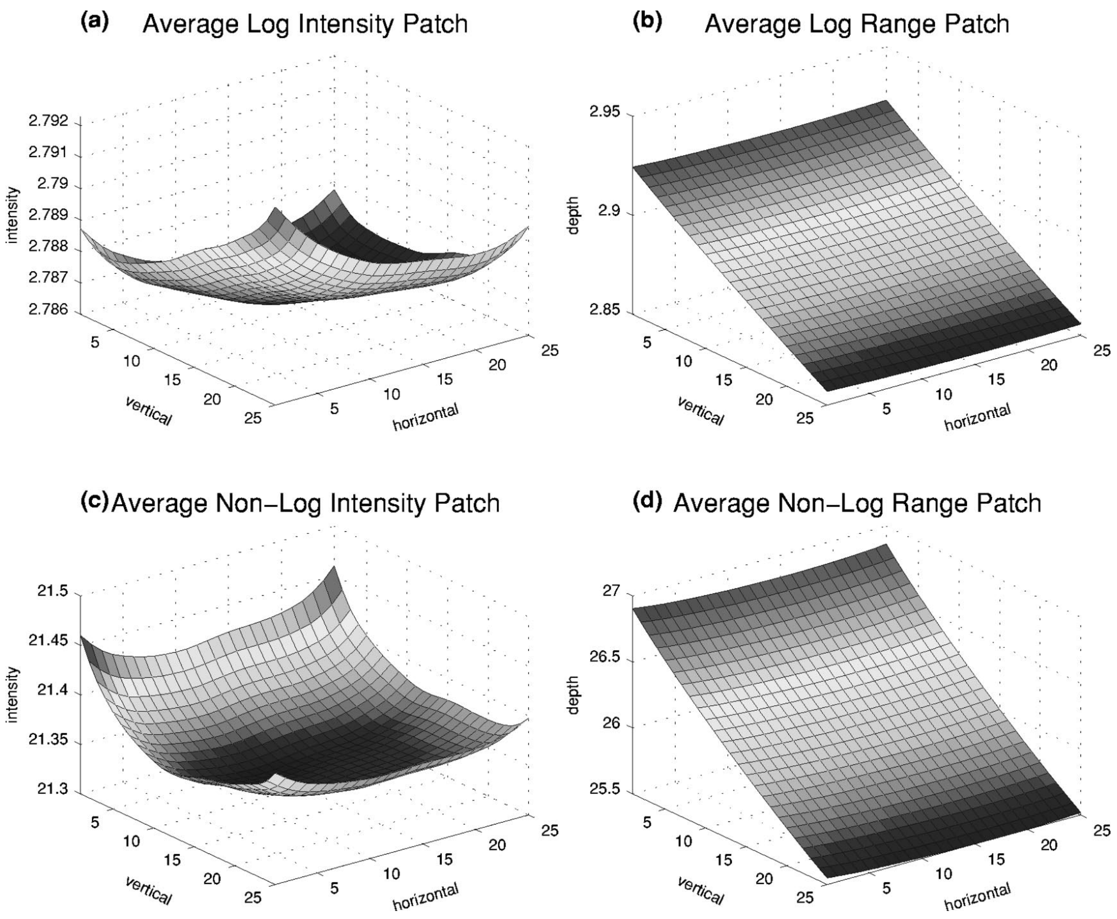

Fig. 2. Averages of all image patches.

multiplicative factor, becomes an additive factor under log intensity. This means that linear filters respond to contrast rather than raw differences in amplitude, and therefore zero-sum linear filters are insensitive to the total contrast of each patch.

We transform the range data in the same way, by applying a logarithmic transform, as was done in previous studies of pure range data. As described by Huang *et al.*,[16] a large object and a small object of the same shape appear identical to the eye when the large object is positioned appropriately far away and the small object is close. However, the raw range measurements of the large, distant object will differ from those of the small object by a constant multiplicative factor. In the log range data, the two objects will differ by an additive constant. Therefore, a zero-sum linear filter will respond identically to the two objects.

In this study, we work with 25 × 25 patches of coregistered light intensity and range data. There are 15,577,472 patches in our database subset that contain no invalid (out-of-range) pixels; 8,872,641 of these patches were from rural images, and 6,704,831 patches were from the urban images. We treat our data as 15.6 million observations of 1250 random variables: 625 luminance variables, and 625 range variables. With a spatial resolution of 22.5 pixels per degree, each patch subtends a visual angle of roughly 1.1 deg.

3. RESULTS

A. Patch Averages

In the study of image statistics, it is often assumed that images are translationally invariant. This assumption is based on the belief that the probability of seeing any image is equal to the probability of seeing the same image, only shifted by some distance and in some direction. However, we do not make this assumption, because some viewing angles might be more likely than others, and the range data might not be translationally invariant. Therefore we must compute the average values of each point of our 25 × 25 image patches independently. Let N be the number of patches in our database, and let L_P and R_P be the Pth log light-intensity patch (luminance) and log range patch, respectively. Then

$$\bar{L}(i, j) = \frac{1}{N} \sum_{P=1}^{N} L_P(i, j), \tag{1}$$

$$\bar{R}(i, j) = \frac{1}{N} \sum_{P=1}^{N} R_P(i, j), \tag{2}$$

where i, j indicates the position within the image patch and $\bar{L}(i, j)$ and $\bar{R}(i, j)$ denote the patch averages.

The mean log intensity and log range image patches are shown in Fig. 2. For explanatory purposes, we also show the mean nonlog intensity and range patches (which were computed before the logarithm was applied). As can be seen in the figure, the average log intensity varies between 2.7863 and 2.7923 across the 25 × 25 patch, for a total range of 0.0060. This is a fairly narrow range for log intensity values, which have a standard deviation of 0.82. This suggests that the mean intensity of luminance images is roughly translationally invariant.

The average range patch, on the other hand, appears to exhibit slant relative to the vertical plane. The average log range values vary between 2.8552 and 2.9244, for a total range of 0.0692. This is a much broader range for log range values, which have a standard deviation of approximately 0.91 log m. Therefore, the structure present in the average log range patch is considerably more significant than the structure present in the average log intensity patch.

To better understand the structure observed in the average range patch, it may be useful to temporarily return to nonlogarithmic range values. Again, the average range patch is very close to planar, with the angle of maximum ascent close to the vertical. This slant arises from the receding ground plane in most range images. Recalling that range values are in meters, this result shows that, on average, within our database, the distance from the observer to the closest object increases at a rate of 1.3 m per visual degree in the vertical direction. We found similar results when we analyzed the rural and urban subsets of the database separately, with little variation across the average luminance patch, and an average incline of 1.5 m/deg in the rural images and 1.0 m/deg in the urban images.

Clearly, this result is highly sensitive to the choice of subject matter for each image and to bias in the orientation of the camera. We do not claim that our results generalize to the set of all real images. An image database of aerial photography, for example, is unlikely to contain a similar statistic. However, images in our database are typical of rural and urban outdoor scenes. We expect similar results to be observed in similar image databases.

B. Covariance

To investigate correlational relationships between luminance and range data, we must understand how the individual pixels in the image patches correlate with one another. We now find the covariance between each pair of pixels in our set of image patches. We begin by centering our data by subtracting the average patch from each patch:

$$\hat{L}_P = L_P - \bar{L}, \tag{3}$$

$$\hat{R}_P = R_P - \bar{R}. \tag{4}$$

Note that translational invariance is not assumed for any data; each of the 1250 variables is centered independently. Let $\mathcal{L}$ and $\mathcal{R}$ be the $N \times 25^2$ matrices of all centered log intensity and log range patches, respectively. Each row of $\mathcal{L}$ and $\mathcal{R}$ is a vectorized image patch, such that

$$\mathcal{L}_{(25m+n),\,p} = \hat{L}_p(n,\, m),$$

where $L_P(0, 0)$ is the upper left of the image patch, and $L_P(24, 0)$ is the upper right of the patch. The 625 × 625 matrix $\mathcal{L}'\mathcal{L}$ contains the covariance between each pair of pixels in the intensity patch. Likewise, $\mathcal{R}'\mathcal{R}$ is the range covariance matrix, and $\mathcal{L}'\mathcal{R}$ is the cross covariance matrix. All of the results reported in the remainder of this paper can be derived from these matrices.

In the top row of Fig. 3 we show the correlation be-

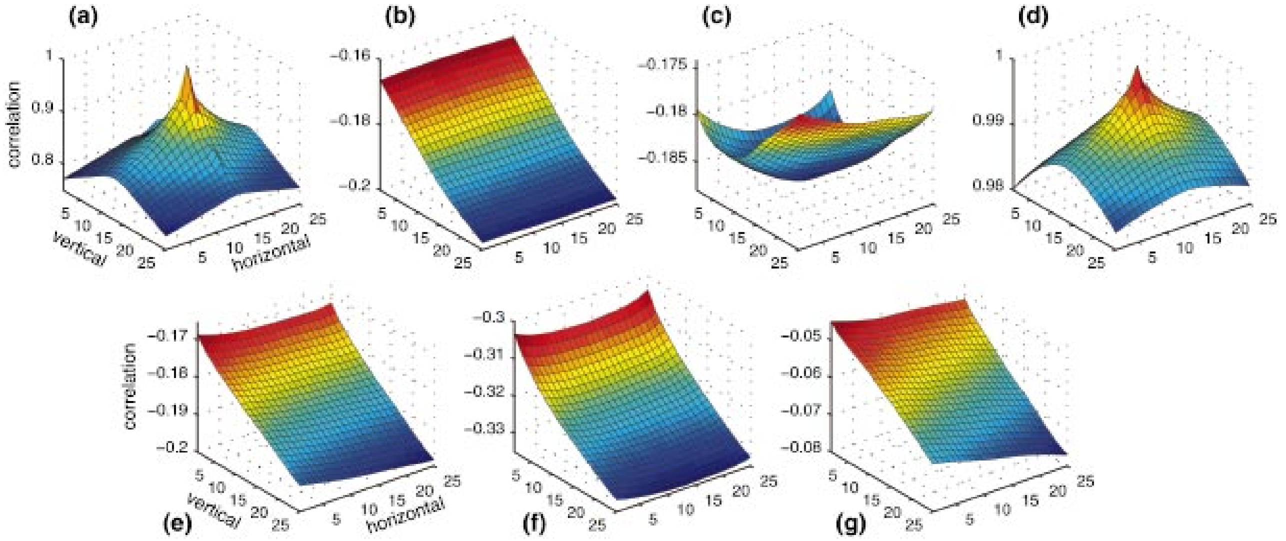

Fig. 3. (a) Correlation between intensity at pixel (13, 13) and all the pixels of the intensity patch. This represents one row of the covariance matrix $\mathcal{L}'\mathcal{L}$, normalized by variance. (b) Correlation between intensity at pixel (13, 13) and the pixels of the range patch. This is one row of the matrix $\mathcal{L}'\mathcal{R}$, normalized by variance. (c) Correlation between range at pixel (13, 13) and the pixels of the intensity patch. This is one column of the matrix $\mathcal{L}'\mathcal{R}$, normalized. (d) Correlation between range at pixel (13, 13) and the pixels of the range patch. This is one row of the matrix $\mathcal{R}'\mathcal{R}$, normalized. (e) Correlation between intensity and range at pixel $(i,\, i)$. This represents the diagonal of the covariance matrix $\mathcal{L}'\mathcal{R}$, normalized by variance. (f) Same figure as (e), except measured over rural images only. (g) Same figure as (e), except measured over urban images only.

tween a specific pixel and other pixels in the patch. The correlation is obtained by means of the equation

$$\rho = \mathrm{cor}[X,Y] = \frac{\mathrm{cov}[X,Y]}{\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}}.$$

We can make several interesting observations from this correlation graph. First, we see that neighboring range pixels are much more highly correlated than neighboring luminance pixels. This suggests that the low-frequency components of range data contain much more power than in luminance images and that the spatial Fourier spectra for range images drops off more quickly than for luminance images, which are known to have roughly $1/f$ spatial Fourier amplitude spectra.[8] Because the measurement of each pixel is independent of the others, any regular distortion of the power spectrum of range images caused by the scanner must be caused by the divergence of the laser beam measuring the distance. For the Riegl LMS-Z360 scanner, the beam divergence is 0.11 degrees, which is approximately 2 pixels in diameter on the images in our database (0.04 to 0.05 deg/pixel). Only the highest spatial frequencies could be influenced by this divergence.

This finding is reasonable because factors that cause high-frequency variation in range images, such as texture or occlusion contours, tend also to cause variation in the luminance image. However, much of the high-frequency variation found in luminance images, such as shadow and surface markings, are not observed in range images.

Second, we see that luminance and range values are negatively correlated, with correlations in the neighborhood of -0.18. Physics-based approaches to shape from shading consistently conclude that shape-from-shading inference offers only relative depth information, not absolute depth information. Our findings suggest that, in natural images brighter pixels have a tendency to be closer to the observer. It has been observed as far back as Leonardo da Vinci that humans perceive brighter objects as closer.[17] Artists have made use of this fact to help create compelling illusions of depth.[18,19] Later, psychologists confirmed this fact in controlled experiments.[20–22] The result presented here is the first evidence that this relationship actually holds in nature.

In psychology literature, this effect is known as relative brightness.[23] Possible explanations are offered as to why such a perceptual bias exists. One common explanation is that light coming from distant objects has a greater tendency to be absorbed by the atmosphere. Under certain weather conditions, such as smog, it is possible for the tendency of the atmosphere to absorb light to outweigh its tendency to scatter light toward the eyes. However, all of the images in our database were taken in clear and sunny conditions. Furthermore, the range of operation for the LMS-Z360 scanner is roughly 200 m, and most of our images were taken under much shorter range. Atmospheric effects are not considered effective until distances considerably further than this.[24]

Another explanation that is given is that nearby objects reflect more light to our eyes.[23] For any given point on an object, the total amount of light that is reflected from that point and into our eye is inversely proportional to the square of the distance to that point. However, the area of the retina occupied by that object also decreases proportionally with the square of the distance to that object. Therefore the amount of light per unit area on the retina remains the same. Except under extreme conditions on a discrete sensor array, this causes perceived brightness to remain the same.

Other explanations are physiological or psychological in nature and thus cannot explain the trend observed in our database. For example, it has been suggested the bias is a combination of the observations that bright objects appear larger (an effect known as irradiance) and larger objects appear closer.[20] Another explanation has been that bright objects are seen more clearly and clear objects are seen as closer, owing to scattering of light in the atmosphere (the aerial perspective).[22] Other theories involve the presence of a third, background stimulus. Studies have shown that against gray or dark backgrounds, brighter stimuli appear closer. However, against a white background, darker stimuli appear closer.[25] Higher order statistical analysis would be needed to search for a physical basis to trends involving the statistical relationship between three stimuli in our database.

As with the patch average results, the correlation found in this database may be dependent on bias in the camera position or orientation. Owing to limitations of the camera, no object in our images is closer than 2 m away, and all of the scenes in our images were brightly lit. The results we obtain might be different if the camera were consistently placed in shaded, cluttered locations, looking out into brightly lit clearings. For example, scenes typical to predators may be different than those typical to prey and therefore may have different statistics.

On inspection of the images in our database, we suspected that the major cause of the correlation was the effects of shadows within the environment. For example, leafy foliage constitutes a significant portion of the database. Since the source of illumination comes from above, and outside any tree, the outermost leaves of a tree or bush are typically the most illuminated. Deeper into the tree, the foliage is more likely to be shadowed by neighboring leaves. On the other hand, urban environments contain more flat surfaces and fewer concavities or places of shadow. We expected the correlation to be much stronger for the rural scenes.

To test this hypothesis, we repeated the analysis on the rural image subset and the urban image subset independently. In the bottom row of Fig. 3 we plot the correlation between range and intensity pixels at each location in the image patch. As can be seen from the figure, the correlations for the rural dataset have strengthened to -0.32, while those for the urban dataset are considerably weaker, in the neighborhood of -0.06.

If the correlation found in the original dataset were due to atmospheric effects, we would expect the correlation to be equally strong in both datasets. The average depth in the urban database (32 m) was similar to that of the rural database (40 m), so atmospheric effects should be similar in both datasets. However, the correlation seems to come almost entirely from the rural scenes.

We feel that the true source of the correlations comes from the statistics of surfaces in natural images and the effects of shadows on these surfaces. In nature, surfaces often have concavities and interiors. Because the light source is typically positioned outside of these concavities, the interiors of these concavities tend to be in shadow and more dimly lit than the object's exterior. At the same time, these concavities will be farther away from the viewer than the object's exterior. Thus the correlation found in an image database should depend a great deal on the statistics of surface shape. In rural images, the leafy structure of foliage and the rocky texture of stone provide an abundance of concavities at any spatial scale. The smooth surfaces of building exteriors offer far fewer opportunities for self-shadowing. It is still possible that significant negative correlation may be observed in databases of indoor scenes. Crumpled fabrics and cluttered environments often contain shadowed concavities.

Langer and Zucker[26] observed that for continuous Lambertian surfaces of constant albedo, lit by a hemisphere of diffuse lighting and viewed from above, a tendency for brighter pixels to be closer to the observer can be predicted from the equations for rendering the scene. Intuitively, the reason for this is that under diffuse lighting conditions the brightest areas of a surface will be those that are the most exposed to the sky. When viewed from above, the peaks of the surface will be closer to the observer. Although these theoretical results have not been extended to more general environments, our results show that in natural scenes, these tendencies remain, even when scenes are viewed from the side, under bright light from one direction. In spite of these differences, both phenomena seem related to the observation that concave areas are more likely to be in shadow.

Finally, we observe that there is structure in the covariance between luminance and range pixels. Not only is the luminance of a pixel correlated with distance at that same pixel, but the same luminance value is even more highly correlated with distance at pixels lower in the patch. This is the first indication of many from our database that bright surfaces tend to face upward. In the remainder of the paper, we present different ways of viewing and understanding these correlations and their structures.

C. Convex Range Filters

In the next three subsections, we will search for those properties in intensity images that correlate most strongly with convexity in range images. We begin by defining a linear filter that models convexity.

We model three types of convexity: vertical convexity, horizontal convexity, and isotropic (nondirectional) convexity. Vertical convexity is modeled as the second vertical derivative of a Gaussian:

$$G(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right), \tag{5}$$

$$\begin{aligned} F_V * R &= \frac{\partial^2}{\partial y^2}[G(x, y) * R(x, y)] \\ &= \frac{\partial^2 G}{\partial y^2} * R, \end{aligned} \tag{6}$$

where R is the range image and σ is set to 6. The second derivative of a function detects the rate of change in the slope of a surface. Recalling that higher range values are more distant, we see that this linear filter will respond strongly for a surface whose tangent plane faces upward toward the top of the patch and downward toward the bottom. This defines vertical convexity. Note that, according to this definition of convexity, corners and points near occlusion boundaries on a foreground object are also considered convex.

Horizontal convexity is modeled similarly, as the second horizontal derivative of a Gaussian. Isotropic convexity is modeled as a Laplacian of Gaussian:

$$F_H * R = \frac{\partial^2 G}{\partial x^2} * R(x, y), \tag{7}$$

$$F_I * R = \nabla^2[G(x, y) * R(x, y)], \tag{8}$$

$$F_I = F_V + F_H. \tag{9}$$

All three filters were normalized to form zero-mean unit vectors:

$$\sum_{i,j} F(i, j) = 0, \tag{10}$$

$$\sum_{i,j} F(i, j)^2 = 1. \tag{11}$$

These three filters are shown in Fig. 4(a).

D. Weighted Averages

Convolving the range filter with the range data provides a measure of surface convexity at each point. We now ask, How do the pixels of a patch in the intensity image covary with this measure of surface convexity? To answer this question, we compute the covariance between the intensity image and the range filter's response. Note that this covariance is equal to the average of all intensity patches, weighted by filter response:

$$\operatorname{cov}[L_P, F \cdot R_P] = \mathcal{L}'\mathcal{R}\mathcal{F}_R \tag{12}$$

$$\begin{aligned} &= \frac{1}{N}\sum_{P=1}^{N} (F \cdot R_P - \mathbf{E}[F \cdot R_P]) \\ &\quad \times (L_P - \mathbf{E}[L_P]) \end{aligned} \tag{13}$$

$$\begin{aligned} &= \frac{1}{N}\sum_{P=1}^{N} (F \cdot R_P - F \cdot \bar{R})(L_P - \bar{L}) \\ &= \frac{1}{N}\sum_{P=1}^{N} (F \cdot \hat{R}_P)\hat{L}_P \\ &= \mathbf{E}[(F \cdot \hat{R}_P)\hat{L}_P], \end{aligned} \tag{14}$$

where $F \cdot R_P$ is the dot product of the linear filter F with the range patch R_P.

These covariances (or weighted averages) are shown in Fig. 4(b). In Fig. 4(c) we show the correlation between the range filter response and each log intensity pixel. The vertical convexity response correlates positively with light intensity toward the top of the patch, and it correlates negatively with light intensity toward the bottom of the patch. This suggests that items that are vertically convex, such as spheres or horizontal cylinders, are more likely to be well lit toward the top of the convexity, whereas concave items are more likely to be bright toward the bottom and dark toward the top.

The weighted average for the horizontal convex filter F_H shows positive correlation with light toward the upper central region of the patch, and this positive correlation extends downward and toward the sides of the patch. One possible explanation for high correlation with intensity toward the top of the patch is that areas of positive Gaussian curvature, such as a sphere, are more common than areas of negative Gaussian curvature, such as the horn of a trumpet. Some studies have suggested that humans have a perceptual bias toward positive Gaussian curvature.[27] It is possible that this bias is matched in nature. If this were the case, horizontally convex regions would tend to be vertically convex as well and would therefore also tend to be brighter toward the top.

E. Linear Correlations

In the preceding subsection we observed that our linear models of convexity correlate with the intensity of light in that region. We now ask three questions: Can we use these correlations to predict the convexity response? What linear intensity filter yields a response that correlates most highly with convexity response? How well do these filter responses correlate?

To answer these questions, we perform linear regression on the range filter response. The regression coefficient F_L is the linear filter that minimizes the sum squared error between the two filter responses:

$$F_L = \arg\min_{F_L} \sum_{P=1}^{N} (F_L \cdot \hat{L}_P - F_R \cdot \hat{R}_P)^2.$$

It can also be shown that the vector F_L maximizes $\mathrm{cor}[F_L \cdot L_P, F_R \cdot R_P]$, just as the vector $\mathcal{L}'\mathcal{R}\mathcal{F}_R$ (the weighted average) lies in the direction of the vector that maximizes $\mathrm{cov}[F_L \cdot L_P, F_R \cdot R_P]$.

Let $\mathcal{F}_R$ be the $25^2 \times 1$ column vector of range filter coefficients for a single linear range filter, and let $\mathcal{F}_L$ be the $25^2 \times 1$ column vector of regression coefficients. Then the regression coefficients are computed as follows[28]:

$$\mathcal{F}_L = (\mathcal{L}'\mathcal{L})^{-1}\mathcal{L}'(\mathcal{R}\mathcal{F}_R). \tag{15}$$

Note that $\mathcal{L}'\mathcal{R}\mathcal{F}_R$ is simply the weighted average, $\mathrm{cov}[L_P, F_R \cdot R_P]$, in vector form, and that $\mathcal{L}'\mathcal{L}$ is simply the covariance matrix of the log intensity values.

The regression coefficients for our three convexity filters are shown in Fig. 5(b). We see that the results are highly noisy. To illustrate why, we perform singular-value decomposition on the symmetric matrix $\mathcal{L}'\mathcal{L} = U'DU$, so that U is an orthonormal matrix whose columns are eigenvectors of $\mathcal{L}'\mathcal{L}$ and D is a diagonal matrix of eigenvalues. We then have

$$\mathcal{F}_L = UD^{-1}U'\mathcal{L}'(\mathcal{R}\mathcal{F}_R), \tag{16}$$

$$U'\mathcal{F}_L = D^{-1}U'\mathcal{L}'(\mathcal{R}\mathcal{F}_R). \tag{17}$$

This shows that when $\mathcal{F}_L$ is projected into the principal components of the luminance images, the coefficients are equal to those of $\mathcal{L}'(\mathcal{R}\mathcal{F}_R)$ divided by the corresponding eigenvalue. This has the effect of amplifying the values of those components with the least variance in the image database, which are therefore the most prone to noise. In the case of natural images, the eigenvectors of lowest eigenvalue are the high-frequency components of the images, and the nth eigenvalue of $\mathcal{L}'\mathcal{L}$ is roughly proportional to $1/n$.[8,15] Thus the high-frequency components

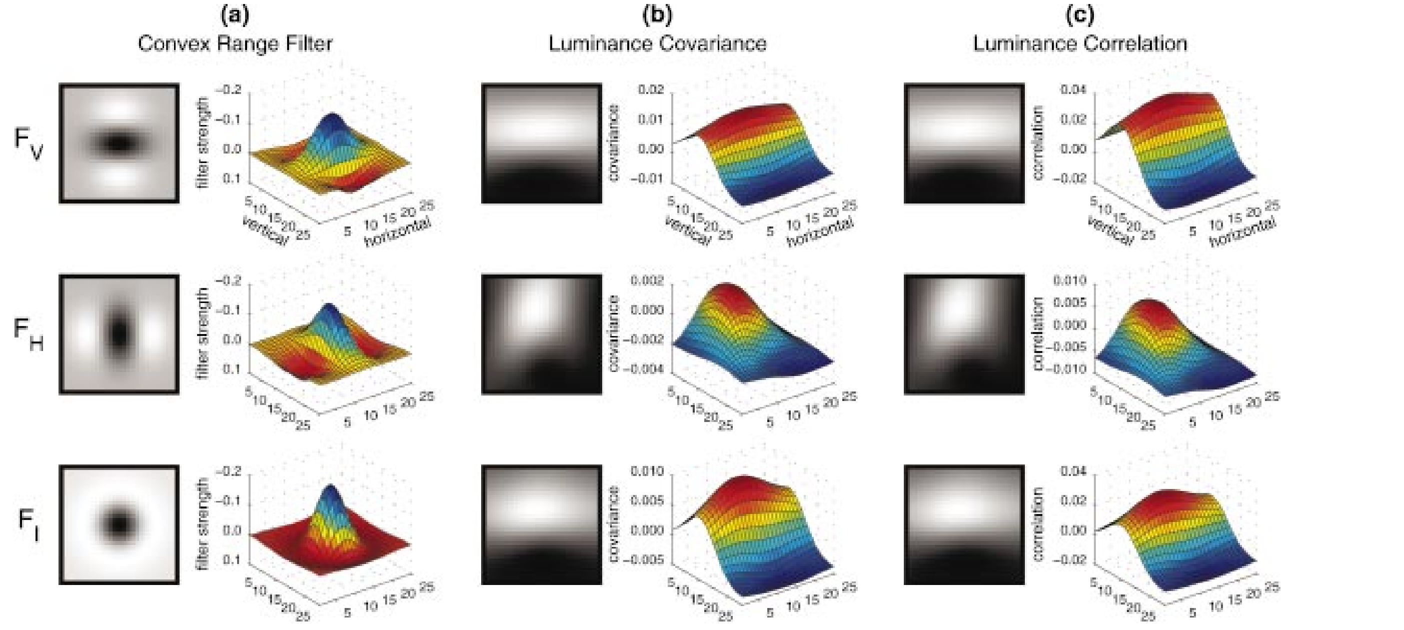

Fig. 4. Convexity weighted averages. (a) Image and surface plots of the three convex range filters, (b) covariance between intensity patch pixels and range filter response, (c) correlation between intensity and range filter response.

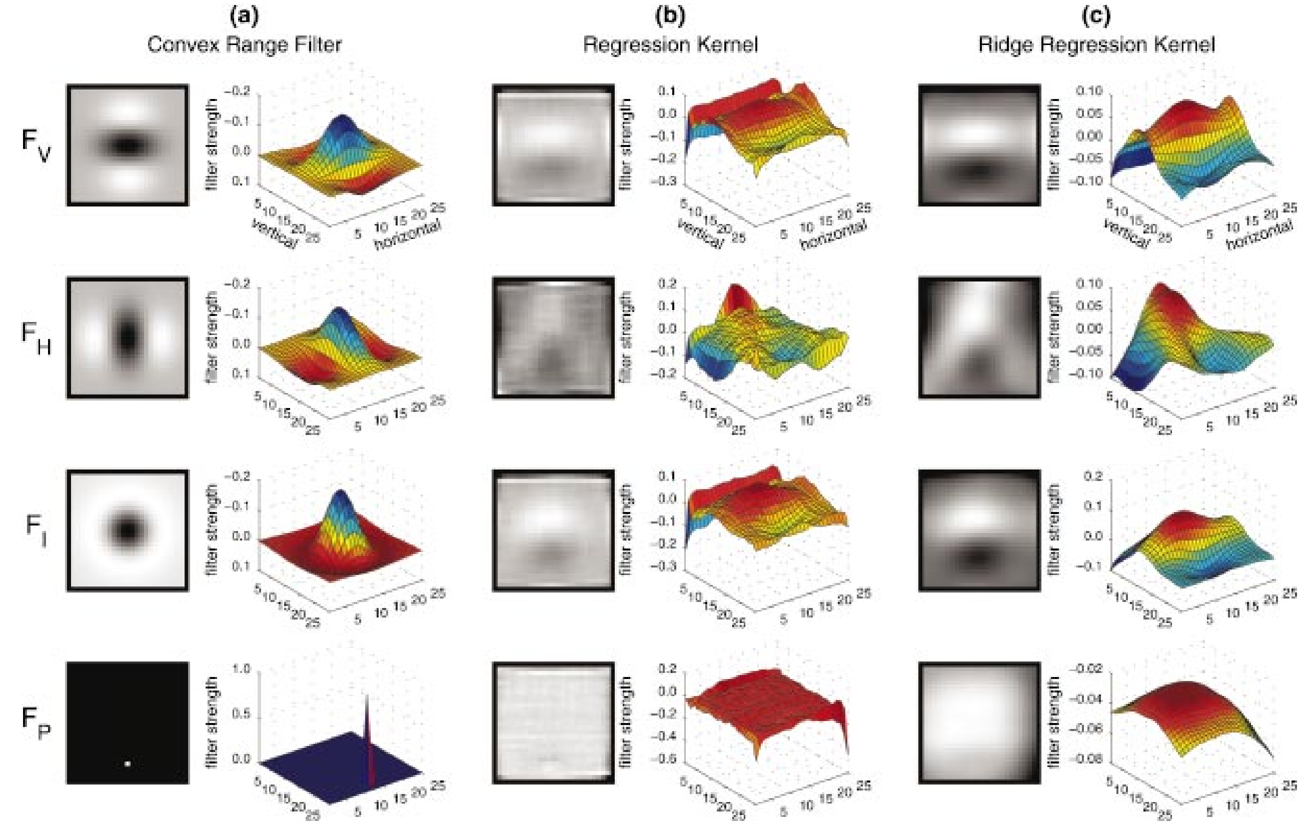

Fig. 5. Linear regression results. Correlation values can be found in Table 1. (a) Convex range filters, (b) ordinary least-squares regression kernels (intensity image filters), (c) ridge regression kernels.

have considerably smaller eigenvalues, and so noise in the high-frequency bands of the weighted averages will be amplified considerably in the regression coefficients. In Fig. 6 we show the coefficients of the principal components of the weighted averages and the regression kernels for our three range filters.

We wish to compensate for the possibility that the large high-frequency coefficients that we observe in the regression kernels are due to noise. We use a standard technique known as ridge regression to simultaneously minimize both sum squared error and the total length of the intensity filter vector $\mathcal{F}_L$. Ridge regression reduces those coefficients that are the least useful in predicting the range filter response,

$$\tilde{F}_L = \arg\min_{F_L}\left[\sum_{P=1}^{N}(F_L \cdot \hat{L}_P - F_R \cdot \hat{R}_P)^2 + \lambda \sum_{i,j} F_L(i,j)^2\right],$$

where $\tilde{F}_L$ is the ridge-regression optical filter, and λ is a parameter of the regression. It can be shown[28] that solving for this equation yields

$$\tilde{\mathcal{F}}_L = (\mathcal{L}'\mathcal{L} + \lambda\mathcal{I})^{-1}\mathcal{L}'(\mathcal{R}\mathcal{F}_R). \tag{18}$$

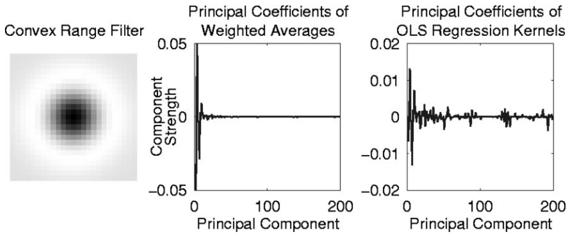

Fig. 6. Principal components coefficients. (a) Convex range filter, (b) principal components coefficients of weighted average, (c) principal components coefficients of ordinary least-squares regression kernels.

For each range filter, we used cross validation to find the value of λ used. For each of the 50 images in our database, we computed the covariance matrices $\mathcal{L}'\mathcal{L}$ and $\mathcal{L}'(\mathcal{R}\mathcal{F}_R)$ from the image subset consisting of each of the 49 remaining images. We then computed $\tilde{F}_L$ for a range of values of λ and tested the sum squared error yielded by each $\tilde{F}_L$ on the one remaining image. λ was chosen to be the value that yielded the smallest mean error of the 50 tests.

The resulting ridge-regression kernels are shown in Fig. 5(c). We see that convexity in range images is best predicted by brighter luminance above the point of convexity, accompanied by darker luminance below. We also see that there is a remarkable resemblance between the regression kernel for isotropic convexity with a Lambertian sphere lit from above.

In addition to showing which properties of intensity images are most correlated with specific linear properties of range images, regression analysis also provides a measure of how strong these relationships are. The values of ρ for each filter are shown in Table 1. The correlational values for the convexity filters are significant, but they are not sufficient to accurately predict convexity from an intensity image in any practical application. However, it should be noted that the original log range image could be fully reconstructed, up to an additive constant, from the response to the isotropic convex filter alone, by using deconvolution. This is because the convex range filters used in this experiment contain all spatial frequencies except very low and very high spatial frequencies. Therefore correlations near 1 would signify that nearly perfect reconstruction of the range image could be achieved from linear processing on the intensity image. Perfect reconstruction is a difficult goal, considering that the images in this database are complex natural images and they contain surfaces that violate the assumptions described in Section 1. Even humans, who could not accurately estimate a complete range image from a natural image, must rely on several monocular and binocular cues to estimate depth information. Modern shape-from-shading techniques typically perform poorly on these images.

To estimate the total predictive power of these simple correlation-based techniques, we use regression to predict a single range pixel from the 25 × 25 intensity patch. The results are shown in Fig. 5 and Table 1. Predicting a single range pixel from each intensity patch provides a way to predict the original range image without introducing noise due to deconvolution. The correlation for this filter is 0.21. It is can be shown from Eq. (15) that

$$\mathrm{cov}[F_L \cdot L_P, F_R \cdot R_P] = \mathrm{var}[F_L \cdot L_P], \tag{19}$$

$$\mathrm{cor}[F_L \cdot L_P, F_R \cdot R_P]^2 = \frac{\mathrm{var}[F_L \cdot L_P]}{\mathrm{var}[F_R \cdot L_R]}. \tag{20}$$

This means that 4% of the total variation in range images can be explained by the intensity image by using simple second-order statistic techniques.

The above calculation estimates how well the intensity image predicts the complete range image and all of its properties. However, we expect that some properties of range images are more easily predicted than others. In the next subsection, we examine those properties of range images that can be best predicted from the luminance image.

F. Canonical Correlation

We have now seen that the responses to convex linear range filters correlate with the responses of certain intensity filters. Our approach so far has been to choose specific range filters and investigate their relationship with coregistered light intensity. In this subsection we extract the filters or features in both domains that are most correlated with or predictive of one another. What we would now like to know is, What linear properties of range images are most easily predicted by the intensity image? Is F_V the range filter whose regression will yield the highest correlation between the filter responses, or are there other range filters that are more correlated with the intensity data?

To answer this question, we use a technique known as canonical correlation. Canonical correlation finds the pair of vectors, F_L^1 and F_R^1 that maximizes the correlation

$$\mathrm{cor}[F_L^1 \cdot L_P, F_R^1 \cdot R_P].$$

It then finds vectors F_L^2 and F_R^2 that maximize correlation subject to the constraint that $\mathrm{cor}[F_L^2 \cdot L_P, F_L^1 \cdot L_P] = 0$ and $\mathrm{cor}[F_R^2 \cdot R_P, F_R^1 \cdot R_P] = 0$. The process repeats until the vectors span the space of image and range patches. It can be shown that F_L^n and F_R^n are proportional to the nth eigenvectors of the matrices $(\mathcal{L}'\mathcal{L})^{-1}(\mathcal{L}'\mathcal{R})(\mathcal{R}'\mathcal{R})^{-1}(\mathcal{R}'\mathcal{L})$ and $(\mathcal{R}'\mathcal{R})^{-1}(\mathcal{R}'\mathcal{L})(\mathcal{L}'\mathcal{L})^{-1} \times (\mathcal{L}'\mathcal{R})$, respectively.[29] Because the correlation we seek to minimize is independent of the magnitude of the vectors F_L^n and F_R^n, we can constrain each filter to have unit variance ($\mathrm{var}[F_L^n \cdot L_P] = 1$) without affecting the value of the correlations. Let $\mathcal{F}_L$ and $\mathcal{F}_R$ denote the $25^2 \times 25^2$ matrices of these eigenvectors. Then, not only do $\mathcal{F}_L$ and $\mathcal{F}_R$ whiten the space of intensity and range patches, but $\mathrm{cor}[F_L^n \cdot L_P, F_R^m \cdot R_P] \neq 0$ implies that $m = n$.

Once again, the resulting linear filters contain a great deal of high-frequency noise. By combining ridge regression with canonical correlation, we obtain a much cleaner result. This technique is known as canonical ridge,[30,31] and it is equivalent to finding eigenvectors of the matrices $(\mathcal{L}'\mathcal{L} + I\lambda_L)^{-1}(\mathcal{L}'\mathcal{R})(\mathcal{R}'\mathcal{R} + I\lambda_R)^{-1}(\mathcal{R}'\mathcal{L})$ and $(\mathcal{R}'\mathcal{R} + I\lambda_R)^{-1}(\mathcal{R}'\mathcal{L})(\mathcal{L}'\mathcal{L} + I\lambda_L)^{-1}(\mathcal{L}'\mathcal{R})$. The results are shown in Fig. 7.

First, it is not surprising that the resulting filters are all highly localized in Fourier space. If the range filters in Fig. 7 had power in all spatial frequencies, then decon-

Table 1. Linear Regression Results for the Three Linear Convexity Filters[a]

Range Filter	λ	ρ	$\tilde{\rho}$	$\tilde{\rho}_R$	$\tilde{\rho}_U$
F_V	0.3	0.1358	0.1353	0.1760	0.1026
F_H	0.7	0.0447	0.0436	0.0744	0.0542
F_I	0.5	0.1053	0.1046	0.1409	0.0862
F_P	10.0	0.2147	0.2142	0.3784	0.0769

[a] ρ is the correlation between the intensity filter response and the ordinary least squares regression range filter response ($\mathrm{cor}[F_L \cdot L_P, F_R \cdot R_P]$). $\tilde{\rho}$ is the correlation after ridge regression. $\tilde{\rho}_R$ and $\tilde{\rho}_U$ are the correlations in the rural and urban datasets, respectively. The fourth filter, F_P is the single-pixel response filter shown in row four of Fig. 5.

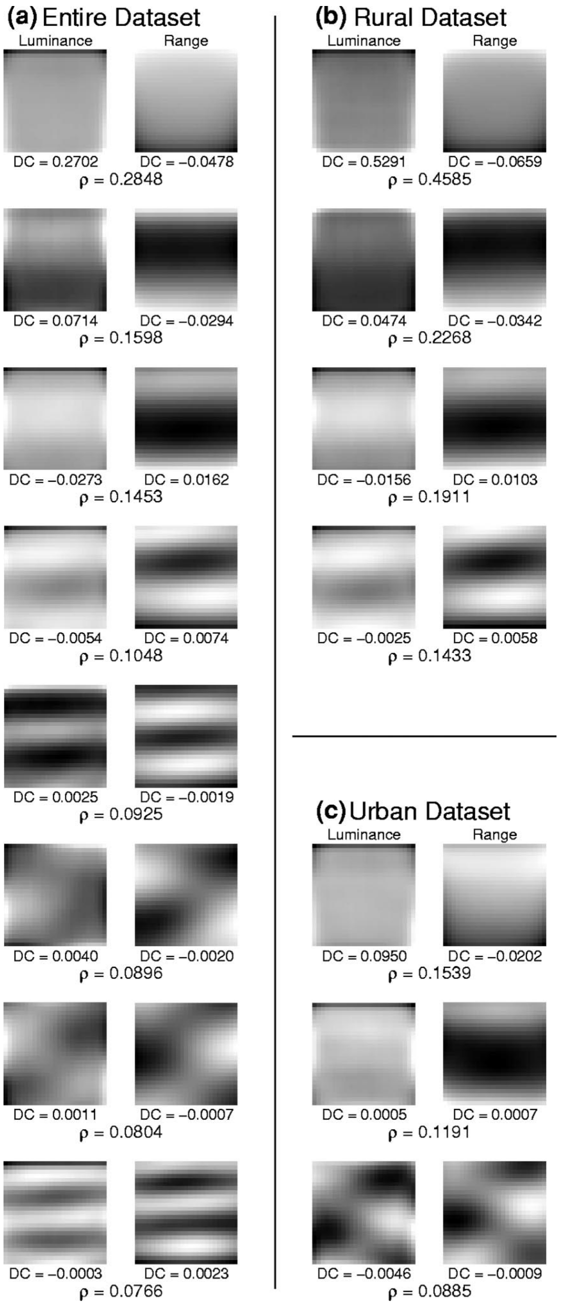

Fig. 7. Canonical correlation filters. (a) First eight canonical ridge filters for the entire (50-image) database. Below each filter is the dc component strength, normalized on a scale from -1 to $+1$. Below each filter pair is listed the correlation strength ρ. (b) First four canonical filters for the rural database. (c) First three canonical filters for the urban database.

volution would allow us to reconstruct the complete range image from prediction on the basis of a single intensity filter. We already have seen that the ability to reconstruct the full range image from correlational statistics is limited.

The next point to observe in these results is that vertical convexity is not the property of range images that is most easily predicted from optical images. The pair of vectors that exhibits the greatest correlation is the dc component of images, which responds strongly to bright image patches, paired with a planar range filter that responds strongly to surfaces with gradients pointing upward and away from the observer. This tells us that the correlation between surfaces facing upward and bright image patches is the highest correlation among all pairs of linear intensity and range filters. The correlation is 0.28. This means that 0.28^2, or 8% of all variation in the overall brightness of a patch of image can be explained by the vertical component of its orientation, and visa versa. In other words, just knowing the overall intensity of an image patch reduces the uncertainty of the vertical slope of that patch by 8%.

The next two canonical variates show vertical convexity. Both of them show that convex surfaces are correlated with bright intensity values that fall somewhere above the area of the surface that is closest to the observer. This trend continues in basis-vector pairs of lower correlation. In general, it appears that many of the first several intensity basis vectors resemble how the corresponding range filter would appear if it were rendered as a surface and lit from above.

Another important observation is that the basis vectors of lower frequency appear earlier within the bases and that vertical frequencies are favored. The canonical correlation bases are independent of the variance of each component, and so this result is not related to the power spectrum of intensity or range images. It means that when linear correlations with intensity data are used, lower-frequency components of the range image are more easily predicted than high-frequency components.

In Figs. 7(b) and 7(c) we show the canonical filters for rural and urban scenes, respectively. Although the rural and urban datasets are disjoint, the first several canonical filters are qualitatively similar between the two datasets. However, the correlations are much stronger in the rural dataset. In fact, throughout this paper, correlations in the rural images have been stronger than those in the urban images. However, it is worth noting that the relative strengths of the correlations is not the same for all properties of range images. The discrepancy between correlations among individual points (bright pixels tend to be closer) is quite high: 0.32 versus 0.06. The discrepancy between the correlations for the first canonical filter is smaller: 0.46 versus 0.15, and the discrepancy between the correlations for the second canonical filter is smaller still: 0.23 versus 0.12. There is a similar discrepancy between the correlations for the convexity filters. This suggests that the trends measured by these correlations (that surfaces tilted upward are brighter and that convex objects are brightest directly above the convexity) are more robust in natural images and less dependent on locale.

Finally, it should be pointed out that the correlation coefficient for the strongest canonical filter measures 0.46 in rural scenes. This means that 21% of all variation in the overall brightness of an image patch in rural images can be explained by the tilt of its underlying surface. This is the strongest statistical trend found in this paper.

4. DISCUSSION

In this paper we have shown that the relationship between the shape of objects and their images depends on statistical trends that cannot be inferred from physical models of light. We have found that this relationship is intimately related to the statistics of lighting directions, the statistics of camera or head orientation, and the statistics of surface shapes in natural scenes. These are sources of information that have not been seriously exploited by traditional shape-from-shading techniques, which instead typically rely on handcrafted environmental priors, such as smoothness and curvature constraints, that have not been verified in natural images. Some of the statistical trends shown here have been suspected by psychologists and artists in the past. In this paper we show the extent to which they hold true in the natural environment, and we explore the structure and causes of these trends. This, and future investigations like this, may help us to understand how these statistical trends might be taken advantage of by computers and by the human brain.

Psychophysical experiments show that humans are able to compute some shape information from shading in a parallel manner.[32] This suggests that the computation is both local and fast. Neurophysiological experiments on awake monkeys also implicated the early visual cortex in the mediation of shape-from-shading pop-out perception.[33] Simple, local statistical relationships between shape and shading like those found in this paper may be exploited by the brain to achieve this level of performance. For example, the image in Fig. 8 gives us an instantaneous perception of convexity and concavity for each stimulus in the array. This may be related to the relatively high degree of correlation between shape convexity and the vertical intensity gradient.

Factoring spaces of high dimensionality is highly useful for performing computations in that space. In this study we used regression and correlation techniques to discover linear bases in the range domain and the luminance domain that are correlated with each other. This image basis provides a factorization of image space that is based on its utility in predicting 3D surface characteristics. Principal components analysis and independent components analysis are useful in image coding and compression. However, these techniques are limited in that they rely only on the statistics of images themselves and do not take into account the relevance of particular properties of images to the computational task at hand. For the solution of any given visual problem, the most important characteristics of an image are not always the ones that vary the greatest, such as the first principal components. In fact, rare events are the often most salient and meaningful. The characteristics in images that are most useful for solving a particular problem depend on the nature of the problem. The codes that we have obtained might be more relevant to the task of inferring 3D structure from 2D information.

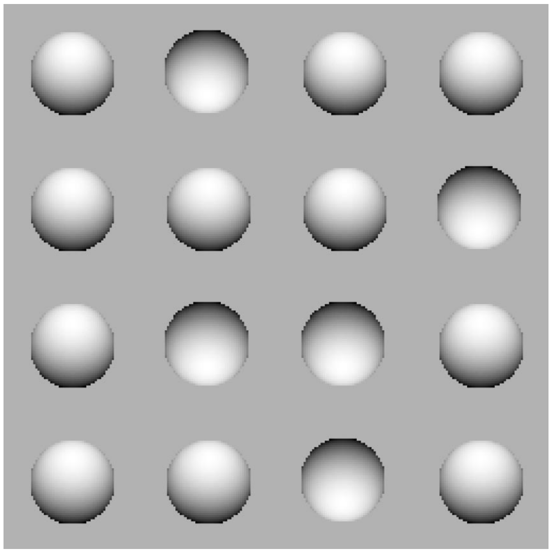
Fig. 8. Convex and concave stimuli defined by shading.

The results in this paper are only the first steps in an investigation into the statistical relationship between shape and shading. We have explored only the simplest relationships between intensity and range images. The methods used in this paper make two strong linearity assumptions. First, the only properties of images considered were their response to linear filters. All previous literature in shape from shading depends on highly nonlinear properties of images. It is possible that stronger results could be achieved by considering such properties. The second assumption of linearity is the measure of correlation between range and intensity image properties. Linear regression finds only properties of range and optical images that are related in a linear fashion. Other measures of relation, such as mutual information, need to be explored. Also, Fig. 3(b) suggests that the correlations measured in this paper occur at spatial distances beyond 25 pixels at 20 pixels/deg. The exact extent of this correlation remains unknown.

ACKNOWLEDGMENTS

We thank Mark Albert, Frankie Li, Kui Shen, and Scott Marmer, who participated in the data collection. We thank Ann Lee and David Mumford for providing some range data for our initial exploration on the range domain. We also thank our reviewers for their insightful comments. Brian Potetz was supported by a National Science Foundation Integrative Graduate Education and Research Traineeships training grant. Tai Sing Lee was supported by National Science Foundation CAREER grant 9984706.

Correspondence may be sent to Brian Potetz at bpotetz@cs.cmu.edu.

REFERENCES

1. B. K. P. Horn, "Obtaining shape from shading information," in *The Psychology of Computer Vision*, P. H. Winston, ed. (McGraw-Hill, New York, 1975), pp. 115–155.
2. W. T. Freeman, E. C. Pasztor, and O. T. Carmicheal, "Learning low-level vision," Int. J. Comput. Vision **40**, 24–57 (2000).
3. S. R. Lehky and T. J. Sejnowski, "Network model for shape-from-shading: neural function arises from both receptive and projective fields," Nature **333**, 452–454 (1988).
4. B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" Vision Res. **37**, 3311–3325 (1997).

5. M. S. Lewicki and B. A. Olshausen, “Probabilistic framework for the adaptation and comparison of image codes,” J. Opt. Soc. Am. A **16**, 1587–1601 (1999).
6. J. H. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” Proc. R. Soc. London Ser. B **265**, 359–366 (1998).
7. P. O. Hoyer and A. Hyvrinen, “Independent component analysis applied to feature extraction from colour and stereo images,” Network Comput. Neural Syst. **11**, 191–210 (2000).
8. D. L. Ruderman and W. Bialek, “Statistics of natural images: scaling in the woods,” Phys. Rev. Lett. **73**, 814–817 (1994).
9. D. J. Field, “Scale-invariance and self-similar ‘wavelet’ transforms: An analysis of natural scenes and mammalian visual systems,” in *Wavelets, Fractals, and Fourier Transforms*, M. Farge, J. C. R. Hunt, and J. C. Vassilicos, eds. (Clarendon, Oxford, UK, 1993), pp. 151–193.
10. J. Huang and D. Mumford, “Statistics of natural images and models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society Press, Los Alamitos, Calif., 1999), pp. 541–547.
11. E. P. Simoncelli, “Modeling the joint statistics of images in the wavelet domain,” in *Wavelet Applications in Signal and Image Processing VII*, M. A. Unser, A. Aldroubi, and A. F. Laine, eds., Proc. SPIE **3813**, 188–195 (1999).
12. C. Q. Howe and D. Purves, “Range image statistics can explain the anomalous perception of length,” Proc. Natl. Acad. Sci. U.S.A. **99**, 13184–13188 (2002).
13. V. S. Ramachandran, “Perception of shape from shading,” Nature **331**, 163–166 (1988).
14. J. Sun and P. Perona, “Preattentive perception of elementary three dimensional shapes,” Vision Res. **36**, 2515–2529 (1996).
15. D. J. Field, “What is the goal of sensory coding?” Neural Comput. **6**, 559–601 (1994).
16. J. Huang, A. B. Lee, and D. Mumford, “Statistics of range images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society Press, Los Alamitos, Calif., 2000), pp. 324–331.
17. E. MacCurdy, ed., *The Notebooks of Leonardo da Vinci, Volume II* (Reynal & Hitchcock, New York, 1938), p. 332.
18. C. Wallschlaeger and C. Busic-Snyder, *Basic Visual Concepts and Principles for Artists, Architects, and Designers* (McGraw Hill, Boston, Mass., 1992).
19. T. M. Sheffield, D. Meyer, J. Lees, B. Payne, E. L. Harvey, M. J. Zeitlin, and G. Kahle, “Geovolume visualization interpretation: a lexicon of basic techniques,” Leading Edge **19**, 518–525 (2000).
20. J. Coules, “Effect of photometric brightness on judgments of distance,” J. Exp. Psychol. **50**, 19–25 (1955).
21. H. Egusa, “Effects of brightness, hue, and saturation on perceived depth between adjacent regions in the visual field,” Perception **12**, 167–175 (1983).
22. I. L. Taylor and F. C. Sumner, “Actual brightness and distance of individual colors when their apparent distance is held constant,” J. Psychol. **19**, 79–85 (1945).
23. D. G. Myers, *Psychology* (Worth, New York, 1995).
24. J. E. Cutting and P. M. Vishton, “Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth,” in *Handbook of Perception and Cognition, Vol 5: Perception of Space and Motion*, W. Epstein and S. Rogers, eds. (Academic, San Diego, Calif., 1995), pp. 69–117.
25. M. Farne, “Brightness as an indicator to distance: relative brightness *per se* or contrast with the background?” Perception **6**, 287–293 (1977).
26. M. S. Langer and S. W. Zucker, “Shape-from-shading on a cloudy day,” J. Opt. Soc. Am. A **11**, 467–478 (1994).
27. P. Mamassian and M. S. Landy, “Observer biases in the 3D interpretation of line drawings,” Vision Res. **38**, 2817–2832 (1998).
28. T. P. Ryan, *Modern Regression Methods* (Wiley-Interscience, New York, 1997).
29. A. Basilevsky, *Statistical Factor Analysis and Related Methods* (Wiley-Interscience, New York, 1994).
30. H. D. Vinod, “Canonical ridge and econometrics of joint production,” J. Econometrics **4**, 147–166 (1976).
31. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Academic, London, 1979).
32. J. Braun, “Shape-from-shading is independent of visual attention and may be a ‘texton’,” Spatial Vision **7**, 311–322 (1993).
33. T. S. Lee, C. Yang, R. D. Romero, and D. Mumford, “Neural activity in early visual cortex reflects behavioral experience and higher order perceptual saliency,” Nat. Neurosci. **5**, 589–597 (2002).