Bayesian depth estimation from monocular natural images

Che-Chun Su Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

Lawrence K. Cormack Department of Psychology, The University of Texas at Austin, Austin, TX, USA

Alan C. Bovik Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

Estimating an accurate and naturalistic dense depth map from a single monocular photographic image is a difficult problem. Nevertheless, human observers have little difficulty understanding the depth structure implied by photographs. Two-dimensional (2D) images of the real-world environment contain significant statistical information regarding the three-dimensional (3D) structure of the world that the vision system likely exploits to compute perceived depth, monocularly as well as binocularly. Toward understanding how this might be accomplished, we propose a Bayesian model of monocular depth computation that recovers detailed 3D scene structures by extracting reliable, robust, depth-sensitive statistical features from single natural images. These features are derived using well-accepted univariate natural scene statistics (NSS) models and recent bivariate/correlation NSS models that describe the relationships between 2D photographic images and their associated depth maps. This is accomplished by building a dictionary of canonical local depth patterns from which NSS features are extracted as prior information. The dictionary is used to create a multivariate Gaussian mixture (MGM) likelihood model that associates local image features with depth patterns. A simple Bayesian predictor is then used to form spatial depth estimates. The depth results produced by the model, despite its simplicity, correlate well with ground-truth depths measured by a current-generation terrestrial light detection and ranging (LIDAR) scanner. Such a strong form of statistical depth information could be used by the visual system when creating overall estimated depth maps incorporating stereopsis, accommodation, and other conditions. Indeed, even in isolation, the Bayesian predictor delivers depth estimates that are competitive with state-of-the-art "computer vision" methods that utilize highly engineered image features and sophisticated machine learning algorithms.

Introduction

By seamlessly combining binocular and monocular cues, humans are able to perceive depths and reconstruct the geometry of the three-dimensional (3D) visual space so quickly and effortlessly that an individual rarely feels how difficult and ill-posed this problem can be. Even given a single color image, or by gazing with one eye closed, a human viewer can still perceive meaningful depth structures and 3D relationships such as relative distances from the visible environment. Yet, automatically estimating range (egocentric distance) from a single monocular image remains a very difficult problem. A variety of factors have been explored to explain how the vision system might accomplish this using depth cues such as color, shading, texture, perspective, and relative size.

A wide variety of computational models have been developed to tackle the problem of depth estimation from a single monocular image. These models typically deploy variants of shape from shading (R. Zhang, Tsai, Cryer, & Shah, 1999; Maki, Watanabe, & Wiles, 2002) and/or shape from texture (Lindeberg & Garding, 1993; Malik & Rosenholtz, 1997). However, the efficacy of these models is typically limited by the information available in luminance and texture variations unless additional structural assumptions or specific constraints are placed on their solutions. Little connection is made with perceptual processes or with the real-world statistical properties that drive them.

Examples of early "computer vision" methods to estimate depths from single images include Hoiem, Efros, and Hebert (2005), who reconstruct a simple 3D model of outdoor scenes assuming that images can be divided into a few planar surfaces, and that pixels can be classified using a small number of limited labels (e.g., ground, sky, and vertical walls). Along similar lines,

Citation: Su, C.-C., Cormack, L. K., & Bovick, A. C. (2017). Bayesian depth estimation from monocular natural images. *Journal of Vision*, *17*(5):22, 1–29, doi:10.1167/17.5.22.

doi: 10.1167/17.5.22 Received March 4, 2016; published May 31, 2017 ISSN 1534-7362

Delage, Lee, and Ng (2006) developed a dynamic Bayesian network to reconstruct the locations of walls, ceilings, and floors by finding the most likely floor–wall boundaries in indoor scenes. Saxena, Sun, and Ng (2009) devised a supervised learning strategy to infer the absolute depth associated with each pixel of a monocular image. They assumed that 3D scenes are made up of small planar surfaces, and used this assumption in conjunction with a Markov random field (MRF) model of textural and luminance gradient cues to infer depth. Nagai, Naruse, Ikehara, and Kurematsu (2002) used hidden Markov models (HMM) to reconstruct surfaces of known classes of objects such as hands and faces from single images. Hassner and Basri (2006) proposed an example-based approach to estimate the depths of objects given a set of known categories. On the other hand, Torralba and Oliva (2002) took a very different (but limited) approach by studying the relationship between the Fourier spectrum of an image and its mean depth. Specifically, they proposed a probability model to estimate the absolute mean depth of a 3D scene using information extracted from the global and local spectral signatures of a 2D image of it. While these methods generally deploy rich sources of low-level information to produce interesting depth estimates, the features used have little connection to known perceptual processes.

More recently, some authors have proposed high-level, ad hoc features to augment single-image depth estimation models. For example, B. Liu, Gould, and Koller (2010) incorporated semantic labels to guide a monocular 3D reconstruction process, thereby achieving better depth estimates. By conditioning on different semantic labels, they were able to better model absolute depth as a function of local pixel appearance. Based on assumed mutual dependencies between semantic labels and geometric depths, Ladický, Shi, and Pollefeys (2014) proposed a joint semantic and unbiased depth classifier that used the property of perspective geometry that the perceived size of an object scales inversely with its distance from the center of projection. Semantic-based algorithms are an interesting computer vision approach, that require considerable manual effort to perform labeling of objects on image and depth training data, and also require subjective decisions regarding what object classes should be defined and labeled. These considerations not only cast doubt on their perceptual relevance, but also may greatly hinder their applicability to practical problems such as vehicular navigation. Karsch, Liu, and Kang (2012) presented an optimization framework to generate a most likely depth map by first matching high-level image features to find candidates from a database, then warping the candidate depth maps under a set of spatial regularization constraints. Utilizing a similar idea of nonparametric sampling, Baig et al. (2014) proposed a depth recovery mechanism by forming a basis over both the RGB and depth feature spaces, and learning a transformation from the RGB to depth weight vectors, where the depth map is estimated as a sparse linear combination of depth-basis elements from the RGB features of the query image. Very recently, deep learning architectures have also been studied for this problem. Eigen, Puhrsch, and Fergus (2014) employ a two-component deep neural network (DNN) to directly regress on depths from single images. The first component estimates the global structure of the scene, while the other refines this estimation locally. While their method delivers promising results, their network architecture is highly engineered and reveals no insights into any perceptual features or processes that might drive single-image depth perception.

Many monocular "shape-from-X" algorithms have also been devised (too many to survey) that estimate relative local depths by assuming the presence of one or more specific attributes, such as texture or shading gradients. Our belief is that such cues are embedded in the local, scale-invariant but space-varying natural statistics of real-world images. Certain natural scene[1] statistics (NSS) models have been shown to provide good descriptions of the statistical laws that govern the behavior of images of the 3D world and 2D images of it. NSS models have proven to be deeply useful tools for both understanding the evolution of human vision systems (HVS; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001) and for modeling diverse visual problems (Portilla, Strela, Wainwright, & Simoncelli, 2003; Tang, Joshi, & Kapoor, 2011; Wang & Bovik, 2011; Bovik, 2013). In particular, there has been work conducted on exploring the 3D NSS of depth/disparity maps of the world, how they correlate with 2D luminance/color NSS, and how such models can be applied. For example, Potetz and Lee (2006) examined the relationships between luminance and range over multiple scales and applied their results to a shape-from-shading problem. Y. Liu, Cormack, and Bovik (2011) explored the statistical relationships between luminance and disparity in the wavelet domain, and applied the derived models to improve a canonical Bayesian stereo algorithm. Su, Cormack, and Bovik (2013) proposed new models of the marginal and conditional statistical distributions of the luminances/chrominances and the disparities/depths associated with natural images, and used these models to significantly improve a chromatic Bayesian stereo algorithm. Recently, Su, Cormack, and Bovik (2014b, 2015a) developed new bivariate and correlation NSS models that capture the dependencies between spatially adjacent bandpass responses like those of area V1 neurons, and applied them to model both natural images and depth maps. The authors further utilized these models to create a blind 3D perceptual image quality model (Su, Cormack, & Bovik, 2015b) that operates on distorted stereoscopic image pairs. An

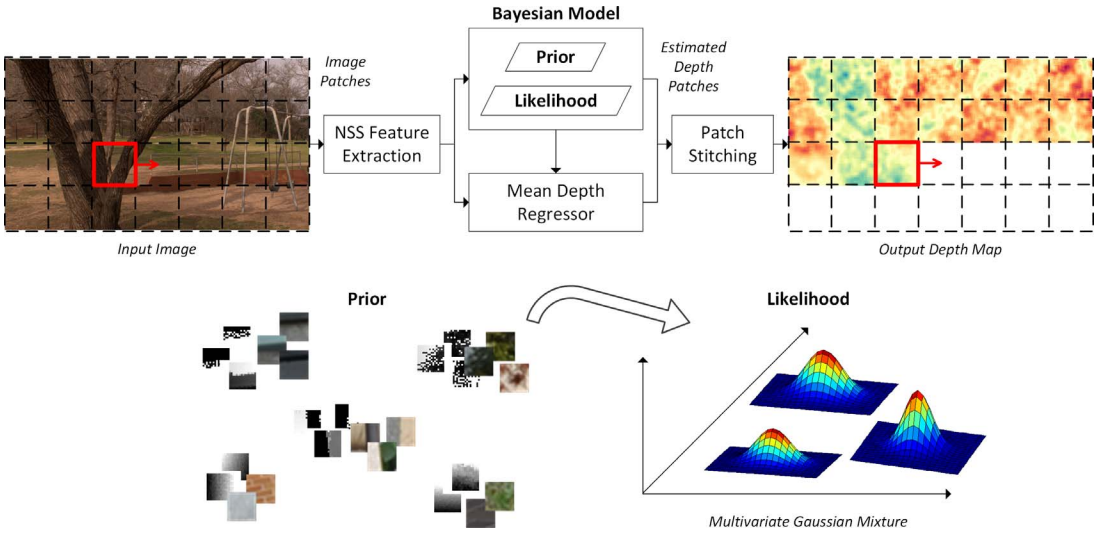


Figure 1. Overview of the proposed Bayesian depth estimation model. The top row depicts the patch-based processing flow of the depth map estimation process on a single natural image, while the bottom row shows the prior and likelihood of the Bayesian model. The priors are a set of representative depth patterns/structures derived from the ground-truth depth patches, while the likelihoods are the conditional probability distributions of the extracted NSS image features given each prior. The red square represents the current estimated patch, while the red arrow points to the next patch to be estimated. For illustrative purposes, the multivariate Gaussian mixture likelihood model is shown in a two-dimensional feature space, instead of the much higher-dimensional feature space defining the depth-from-luminance model. See text for more details.

algorithm derived from this model was shown to deliver quality predictions that correlate very highly with recorded human subjective judgments of 3D picture quality.

Inspired by these successes, and by psychophysical evidence of NSS-driven signal processing in the vision system (Field, 1987, 1999), we describe a Bayesian model for estimating depths from single monocular images that employs reliable and robust NSS models of natural images and depth maps as priors (Su et al., 2014b, 2015a). We trained and tested the model on three publicly accessible databases of natural image and range data: the LIVE 3D+Color Database Release-2 (Su, Cormack, & Bovik, 2016b), which consists of 99 pairs of high-definition resolution (1920 × 1080 pixel) natural images and accurately coregistered high-definition ground-truth depth maps, the Make3D Laser+Image Dataset-1 (Saxena, Chung, & Ng, 2005; Saxena, Sun, & Ng, 2005; Saxena et al., 2009), and the NYU Depth Dataset V2 (Silberman, Hoiem, Kohli, & Fergus, 2012; Silberman, Kohli, Hoiem, & Fergus, 2012).

Proposed Bayesian depth estimation model

We begin by summarizing our Bayesian depth estimation model and the contributions we make. Figure 1 is an overview of the proposed model, in which the top row depicts the flow of the process of depth map estimation from an image, while the bottom row illustrates the prior and likelihood of the Bayesian model. Our depth estimation model is patch-based: an input image is divided into patches of size[2] $P \times P$, a set of perceptual NSS features is extracted from each image patch, then a Bayesian model uses the extracted NSS image feature to form an estimate of the pattern/structure of the corresponding depth patch, offset by a regressed mean depth value; finally, all estimated depth patches are stitched together to create the output depth map. The priors are a set of representative depth patterns/structures derived from the ground-truth depth patches, and the likelihoods are the conditional probability distributions of the extracted NSS image features given each prior. As illustrated in the bottom row of Figure 1, an estimate of a depth patch pattern/structure is calculated as the cluster centroid of the most probable prior given the extracted NSS image feature. The Bayesian prior and likelihood models, as well as the mean depth regressor, are learned from perceptually relevant features extracted from a high-quality 3D database of natural images and accurately coregistered depth maps. These NSS models capture valuable statistical relationships that are embedded in the luminances and depths of the natural images, making reliable depth estimation from monocular natural images possible. The details of each component, including how the perpetual NSS feature set is extracted from each image patch, how the prior and likelihood models are constructed, and how the mean

depth regressor is trained, are explained in the following subsections.

The contributions we make are as follows. By employing established univariate and new bivariate/correlation NSS models of natural images, we define a set of depth-sensitive features representative of perceptually available depth information. We cast the depth recovery problem as a Bayesian inference that is solved in a single step without the need for any assumed high-level semantics, smoothness constraints, or iterative optimization methods. Toward validating the method on a science-quality set of data, we created a high-quality 3D database of high-resolution naturalistic stereoscopic color image pairs with accurately coregistered dense depth maps obtained by a precision LIDAR terrestrial range scanner. This new and accurate database provides a rich source of information on which natural depth statistics can be computed. We are making this database publicly available free of charge (Su et al., 2016b). Despite its simplicity, the Bayesian depth estimation model that we construct using simple natural-scene statistic priors delivers performance that is highly competitive with and even exceeds that of top-performing state-of-the-art depth estimation algorithms that deploy sophisticated deep learning architectures or highly engineered image heuristics. We are also making available the code of our simple NSS-based depth estimation model for independent evaluation and further academic research (Su, 2016).

Perceptual decomposition

Human vision systems (HVS) extract abundant information from natural environments by processing visual stimuli through massively parallel and pipelined levels of decomposition and interpretation. By analyzing the natural statistics of the 2D and 3D visual world, and by learning how the HVS processes natural image and depth information, a variety of statistical models have been proposed that capture the behavior of perceptually motivated bandpass responses of luminance/chrominance and depth/disparity on natural scenes (Field, 1987; Ruderman, 1994; Wainwright, Schwartz, & Simoncelli, 2002; Potetz & Lee, 2003; Y. Liu et al., 2011; Su et al., 2013). Since the philosophy underlying our approach is to learn and employ good models of the statistical laws that describe the relationships between depth perception and the structure of natural images, we apply certain perceptually relevant preprocessing steps to the recorded image data, including biologically motivated linear bandpass decompositions and nonlinear divisive normalization processes. A set of depth-sensitive NSS image features are then extracted from the univariate and bivariate empirical distributions of these responses.

Our work is therefore perceptually motivated and, hence, could be particularly applicable to problems in perceptual image engineering, such as creating 3D presentations suitable for human viewing, as for example in the creation of 3D cinematic or television content from archived 2D movies. However, as we show in the sequel, the method delivers highly competitive objective results, and we envision that, given its conceptual and computational simplicity, it could find other (e.g., robotic) applications. In its current form, we utilize only luminance information in our model, although there are definite statistical relationships between image color and depth (Su et al., 2013), as well as on the perception of depth on color images (Jordan, Geisler, & Bovik, 1990).

We acquire luminance from color images by transforming them into the perceptually uniform CIELAB color space (X. Zhang & Wandell, 1997; Rajashekar, Wang, & Simoncelli, 2010). Each luminance image (L*) is then decomposed by a steerable pyramid decomposition, which is an overcomplete wavelet transform that allows for increased orientation selectivity (Simoncelli & Freeman, 1995; Portilla & Simoncelli, 2000). The use of the wavelet transform is motivated by the fact that its space-scale–orientation decomposition is similar to the bandpass filtering that occurs in area V1 of primary visual cortex (Field, 1987; Olshausen & Field, 2005). It is also a computationally efficient and convenient decomposition. Similar to a conventional orthogonal wavelet decomposition, the steerable pyramid recursively splits an image into a set of oriented subbands and a low-pass residual band (Portilla & Simoncelli, 2000). Figure 2 shows a block diagram of the steerable pyramid decomposition, in which each block represents a filter in the 2D Fourier domain. Specifically, the filters involved in the decomposition are polar-separable in the 2D Fourier domain, and by using the polar frequency coordinate $(r_\omega, \theta_\omega)$, where $\omega = \begin{bmatrix}\omega_x \omega_y\end{bmatrix}^\top = [r_\omega \cos\theta_\omega r_\omega \sin\theta_\omega]^\top$, they can be written as:

$$L(r_\omega, \theta_\omega) = \begin{cases} 2, & r_\omega \le \frac{\pi}{4} \\ 2\cos\left(\frac{\pi}{2}\log_2\left(\frac{4r_\omega}{\pi}\right)\right), & \frac{\pi}{4} < r_\omega < \frac{\pi}{2} \\ 0, & r_\omega \ge \frac{\pi}{2} \end{cases} \quad (1)$$

and

$$B_k(r_\omega, \theta_\omega) = H(r_\omega) G_k(\theta_\omega) \quad (2)$$

where $\in [0, K-1]$, and K is the number of orientations. The radial and angular parts, $H(r_\omega)$ and $G_k(\theta_\omega)$, can be written as:

$$H(r_\omega) = \begin{cases} 0, & r_\omega \le \frac{\pi}{4} \\ \cos\left(\frac{\pi}{2}\log_2\left(\frac{2r_\omega}{\pi}\right)\right), & \frac{\pi}{4} < r_\omega < \frac{\pi}{2} \\ 1, & r_\omega \ge \frac{\pi}{2} \end{cases} \quad (3)$$

and

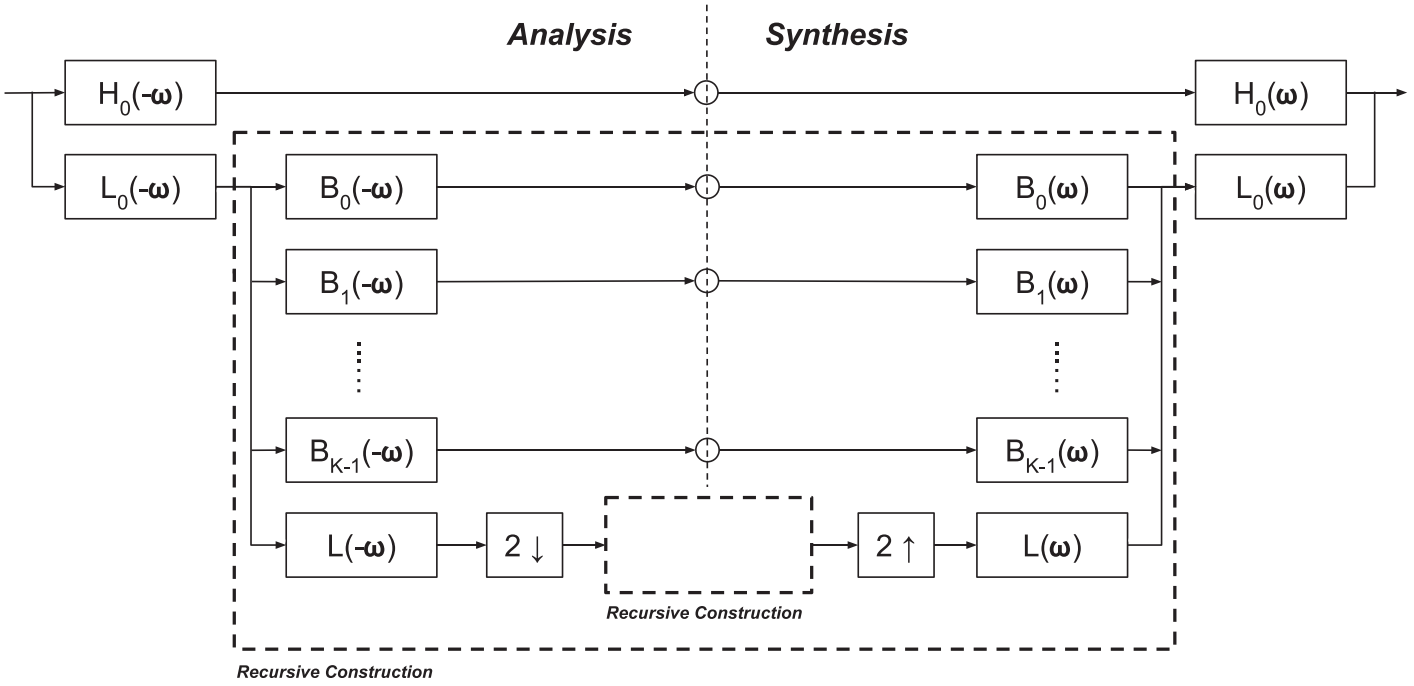


Figure 2. Block diagram of a steerable pyramid decomposition, including both analysis and synthesis filter banks. The input image is split into high- and low-pass bands, and the low-pass band is further split into a set of oriented subbands and another low-pass band. The recursive decomposition takes a down-sampled low-pass band and repeats the same subband decomposition at the next (coarser) scale. From Portilla and Simoncelli (2000).

$$G_k(\theta_\omega) = \begin{cases} b_k\left[\cos\left(\theta_\omega - \frac{\pi k}{K}\right)\right]^{K-1}, & \left|\theta_\omega - \frac{\pi k}{K}\right| < \frac{\pi}{2} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $b_k = 2^{k-1}\frac{(K-1)!}{\sqrt{K[2(K-1)]!}}$. Finally, the steerable pyramid decomposition is initialized by splitting the image into high-pass and low-pass parts using the two filters:

$$L_0(r_\omega, \theta_\omega) = \frac{L\left(\frac{r_\omega}{2}, \theta_\omega\right)}{2} \quad (5)$$

and

$$H_0(r_\omega, \theta_\omega) = H\left(\frac{r_\omega}{2}, \theta_\omega\right) \quad (6)$$

Interested readers may refer to Portilla and Simoncelli (2000) for further details about the steerable pyramid decomposition. In our implementation, we deployed a steerable pyramid decomposition and used the responses of the filters over the two finest scales, and over four canonical orientations: $0, \frac{1}{4}\pi, \frac{1}{2}\pi$, and $\frac{3}{4}\pi$ (rad). We computed NSS features from those subband coefficients and used them in the depth estimation process. Therefore, a total of 2 (scales) × 4 (orientations) = 8 subband responses are computed on each image patch.

After applying the multiscale, multiorientation decomposition, we perform the perceptually significant process of divisive normalization on the luminance wavelet coefficients of all of the subbands (Wainwright et al., 2002). Divisive normalization, or adaptive gain control, accounts for the nonlinear behavior of cortical neurons (Heeger, 1992; O. Schwartz & Simoncelli, 2001). In Simoncelli (1999) and Wainwright et al. (2002), the authors found that the coefficients of orthonormal wavelet decompositions of natural images are fairly well decorrelated; however, they are not independent. The authors also showed that the empirical joint histograms of adjacent coefficients produce contour plots having distinct "bowtie" shapes, which were observed on coefficient pairs separated by different spatial offsets, across adjacent scales, and at orthogonal orientations. These findings suggest that different types of divisive normalization processes over, for example, neighborhoods of scale, orientation, and spatial location occur in primary visual cortex.

The divisive normalization transform (DNT) that we use is Lyu (2011):

$$t(x_i, y_i; s, r) = \frac{v(x_i, y_i; s, r)}{\sqrt{a + \sum_{|x_j - x_i| \le k, |y_j - y_i| \le k} g(x_j, y_j)\, v(x_j, y_j; s, r)^2}} \quad (7)$$

where (x_i, y_i) are spatial coordinates; s and r denote subband scale and orientation, respectively; v are the

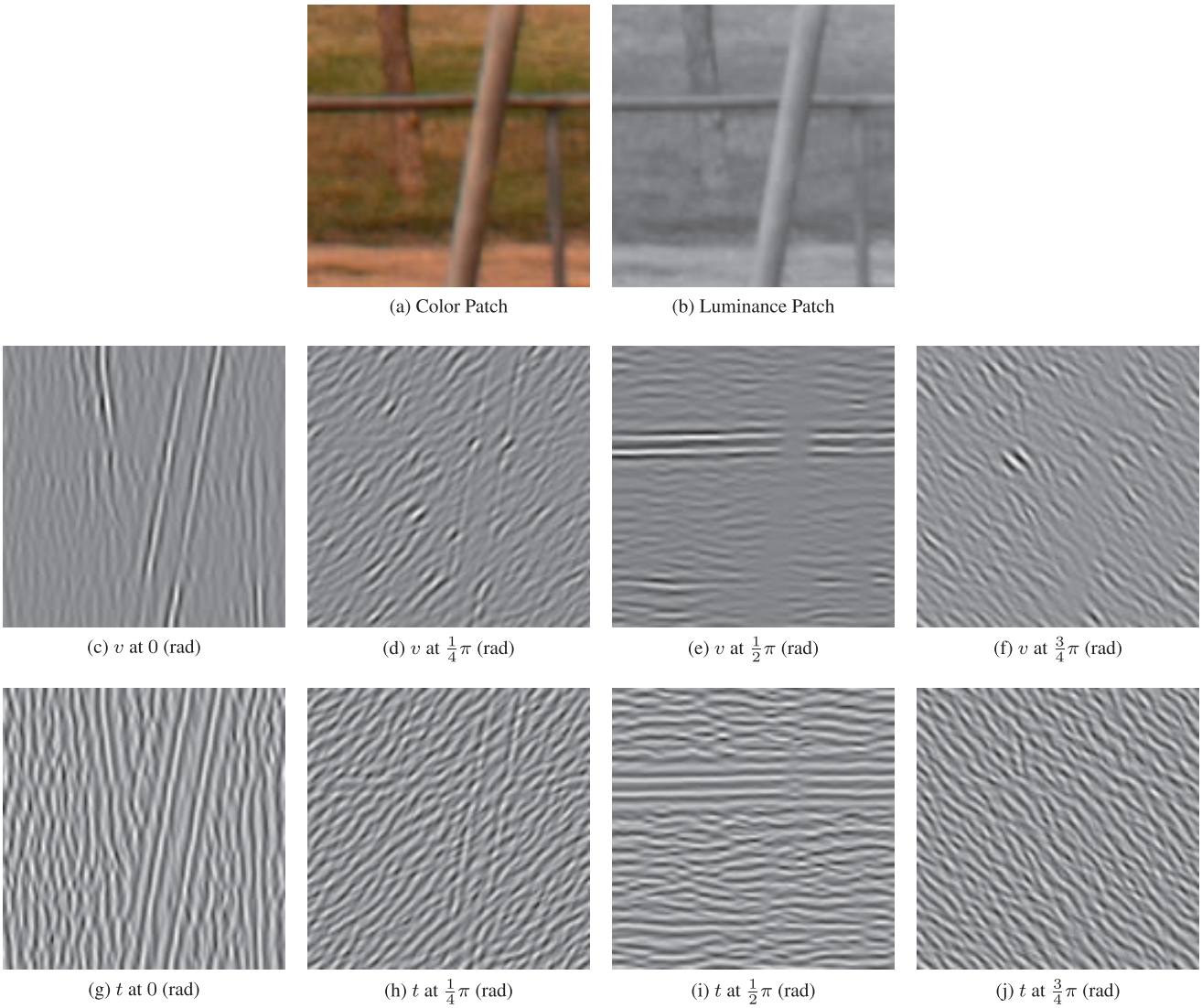


Figure 3. Perceptual decomposition of an example patch selected from the image in Figure 1. Top row: the original color patch and the luminance patch. Middle row: the steerable pyramid subband responses from four canonical orientations at the finest scale. Bottom row: the same subband responses after DNT. The example patch size is 128 × 128 ($P = 128$).

subband coefficients; t are the coefficients following the DNT; a is a semisaturation constant; and $\{g(x_j, y_j)\}$ is a finite-extent Gaussian weighting function with a window size equal to k and $\sigma = k/2$:

$$g(x_j, y_j) = \begin{cases} Ce^{-\frac{(x_j - x_i)^2 + (y_j - y_i)^2}{2\sigma^2}}, & |x_j - x_i| \le k \text{ and } |y_j - y_i| \le k \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where C is a constant that makes $\{g(x_j, y_j)\}$ sum to 1. We perform the DNT for each wavelet coefficient across spatial neighborhoods using a fixed Gaussian window (e.g., 5 × 5 regardless of scale, within each subband. Among the different types of divisive normalization processes mentioned above, we deploy the aforementioned spatial DNT in our implementation due to its simplicity and effectiveness. We found no significant difference in depth estimation performance using more complicated DNTs in our experiments.

Figure 3 shows an example patch selected from the input image in Figure 1, along with its perceptual decomposition using the steerable pyramid and DNT described above. Note that in our implementation, the wavelet decomposition and DNT are performed on the entire input image, from which each patch is cropped to estimate the corresponding depth patch. Zeros are used when pixel coordinates are outside of the image boundary. Different orientation subbands generate larger responses along the corresponding texture/structure directions in the image patch, while the DNT further normalizes the subband responses using the spatial neighborhoods. Next, we introduce the three NSS models (univariate, bivariate, and correlation models)

used by our Bayesian depth estimator, and we explain how we utilize these models to extract the depth-sensitive NSS features from the perceptually processed subband responses (i.e., the wavelet coefficients subjected to the DNT).

Image feature extraction

It is well established that there exist statistical relationships between image luminances and depth information in natural scenes (Potetz & Lee, 2003), and a variety of univariate statistical models have been proposed to fit the bandpass responses of luminance/chrominance and disparity (Y. Liu et al., 2011; Su et al., 2013). Very recently, new closed-form bivariate and correlation statistical models have been developed that effectively capture spatial dependencies between neighboring subband responses in natural images (Su et al., 2014b, 2015a). The Bayesian model of depth estimation we develop here exploits these NSS features to learn the relationships that exist between projected image luminances and colocated depth information. These NSS features are extracted from each subband of the perceptually decomposed image patch as described in the previous subsection; therefore, the number of feature dimensions of each image patch is equal to the sum of the number of parameters of each NSS model times the number of subbands upon which each NSS model is built. In the following subsections, we explain in detail the univariate, bivariate, and correlation NSS models that drive our Bayesian depth estimator.

Univariate NSS model

Considerable work (Field, 1999; Simoncelli & Olshausen, 2001) has been conducted on modeling the statistics of natural images that have been passed through multiscale, multiorientation bandpass transforms (e.g., decompositions using banks of Gabor filters or wavelets). A common and well-accepted model of the empirical histograms of divisively normalized luminance subband responses (i.e., t in Equation 7), is the univariate generalized Gaussian distribution (GGD; Mallat, 1989a, 1989b; Li & Wang, 2009). The probability density function of a univariate GGD with zero mean is:

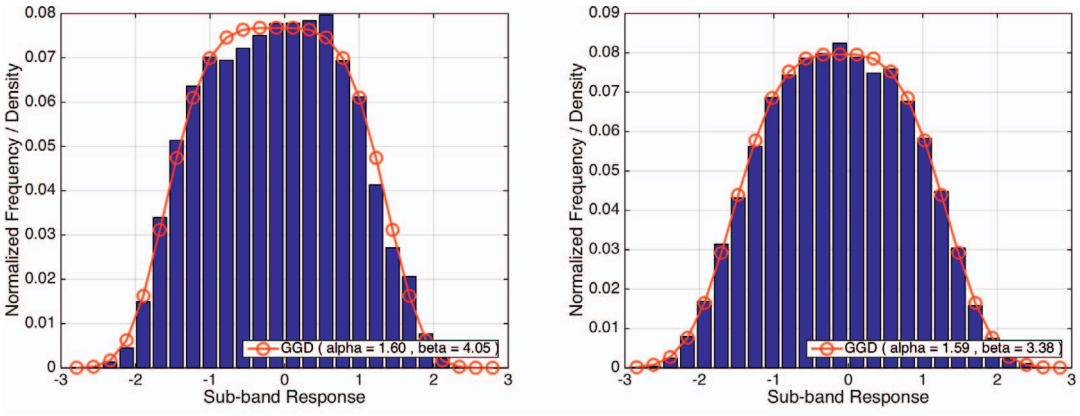
$$p(x;\alpha_u,\beta_u)=\frac{\beta_u}{2\alpha_u\Gamma\left(\frac{1}{\beta_u}\right)}e^{-\left(\frac{|x|}{\alpha_u}\right)^{\beta_u}} \quad (9)$$

where $\Gamma(\cdot)$ is the ordinary gamma function and α_u and β_u are scale and shape parameters, respectively. Note that both v and t in Equation 7 are zero-mean coefficients, since the steerable filters have zero DC response (which is a necessary condition that the transform be invertible) unlike, for example, the filters used in ad hoc Gabor filter decompositions (Clark & Bovik, 1989). Hence, we model t as x in Equation 9. For pristine, undistorted natural images, it is commonly assumed that $\alpha_u \approx \sqrt{2}$ and β_u is close to 2 (i.e., unit-variance Gaussian), while distortions tend to create structural degradations that modify α_u and β_u; typically β_u is closer to and often less than one on distorted images (Sheikh & Bovik, 2006; Moorthy & Bovik, 2011). For example, if an image is blurred, the sparse high-frequency subband responses are eliminated, producing an even larger preponderance of low-frequency responses. The result is a peakier distribution. Thus, an objective of reconstructing a naturalistic image of luminance or depth is that the reconstruction approximately satisfies this Gaussian assumption. Therefore, these parameters may be viewed as constraints on the "naturalness" of the reconstruction. We estimate the GGD parameters on small $P \times P$ patches, so β_u locally varies. The two GGD parameters, α_u and β_u, for all eight subbands are estimated from each subband patch histogram (using the widely used maximum likelihood method described in Sharifi & Leon-Garcia, 1995). These simple measurements are the first elements, 16 numbers, 2 (α_u and β_u) $\times$ 8 (2 scales with 4 orientations each), of the feature set that we define on each image patch.

Figure 4 shows the univariate GGD models fitting the empirical histogram of the subband responses from the example image patch in Figure 3. It can be seen that the empirical histograms from different subband tuning orientations have different shapes, and that the corresponding univariate GGD fits are able to capture these distinct characteristics using the model parameters α_u and β_u.

Bivariate NSS model

We also capture dependencies that exist between spatially neighboring luminance subband responses by modeling the bivariate distributions of horizontally adjacent subband responses sampled from all locations, (x, y) and $(x+1, y)$, of each orientation subband at different scales on each image patch. Since we have observed similar statistics from both horizontally and vertically neighboring responses (Su et al., 2014b, 2015a), and used subband orientations covering 0 to π (rad), we exploit only horizontal adjacency to achieve the same efficacy with reduced computational complexity. This also applies to the correlation NSS feature, which will be detailed in the next section. To model these empirical joint histograms, we utilize a multivariate generalized Gaussian distribution (MGGD), which includes both the multivariate Gaussian and Laplacian distributions as special cases. The probability density function of an MGGD is defined as:



(a) 0 (rad) (b) $\frac{1}{4}\pi$ (rad)

(c) $\frac{1}{2}\pi$ (rad) (d) $\frac{3}{4}\pi$ (rad)

Figure 4. Best-fitting univariate GGD models to the perceptually decomposed subband responses, *t*, using the same example image patch as in Figure 3. The blue bars represent the empirical histograms of the subband responses, while the red circle lines are the fitted univariate GGD models.

$$p(\mathbf{x};\mathbf{M},\alpha_b,\beta_b)=\frac{1}{|\mathbf{M}|^{\frac{1}{2}}}g_{\alpha_b,\beta_b}\left(\mathbf{x}^\top\mathbf{M}^{-1}\mathbf{x}\right) \quad (10)$$

where $\mathbf{x}\in\mathbb{R}^N$, $\mathbf{M}$ is an $N\times N$ symmetric scatter matrix, α_b and β_b are scale and shape parameters, respectively, and $g_{\alpha_b,\beta_b}(\cdot)$ is the density generator:

$$g_{\alpha_b,\beta_b}(y)=\frac{\beta_b\Gamma\left(\frac{N}{2}\right)}{\left(2^{\frac{1}{\beta_b}}\pi\alpha_b\right)^{\frac{N}{2}}\Gamma\left(\frac{N}{2\beta_b}\right)}e^{-\frac{1}{2}\left(\frac{y}{\alpha_b}\right)^{\beta_b}} \quad (11)$$

where $y\in\mathbb{R}^+$. Note that when $\beta_b=0.5$, Equation 10 becomes the multivariate Laplacian distribution, and when $\beta_b=1$, Equation 10 corresponds to the multivariate Gaussian distribution. When $\beta_b\to\infty$, the MGGD converges to a multivariate uniform distribution, and when $\beta_b<0.5$, it becomes a 2D heavy-tailed "sparsity" density. The scatter matrix $\mathbf{M}$ is a sample statistic that can be used to estimate the covariance matrix of $\mathbf{x}\in\mathbb{R}^N$, which may embed dependencies in $\mathbf{x}\in\mathbb{R}^N$ (i.e., the spatially neighboring bandpass image responses). In order to capture these second-order statistics, we adopt a closed-form correlation model, which is described in detail in the next subsection, to extract the corresponding NSS features. In our implementation, we model the bivariate empirical histograms of horizontally adjacent subband responses of each image patch using a bivariate generalized Gaussian distribution (BGGD) with $N=2$ in Equation 10. Specifically, from each subband of an image patch, we collect all pairs of horizontally adjacent subband responses to form $\mathbf{x}\in\mathbb{R}^2$, and estimate the BGGD model parameters using the maximum likelihood estimator (MLE) algorithm described in (Su, Cormack, & Bovik, 2014a). In our case, the scatter matrix $\mathbf{M}$ is a 2×2 matrix, which can be written as:

$$\mathbf{M}=\sum_{j=1}^{n}\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)^\top,$$

$$\text{where } \bar{\mathbf{x}}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j \quad (12)$$

and $n = P \times (P-1)$ is the number of horizontally adjacent pairs in an image patch of size $P \times P$. Both the BGGD scale and shape parameters, α_b and β_b, from all eight subbands are included in each image patch's feature set.

To demonstrate the effectiveness of the BGGD model and the embedded dependencies between horizontally adjacent subband responses, Figure 5 shows 3D plots and 2D isoprobability contours of the bivariate empirical histograms of horizontally adjacent responses, and the corresponding BGGD fits at different subband tuning orientations from the example image patch in Figure 3. As can be seen in both 3D illustrations, where the blue bars represent the empirical histograms and the colored meshes represent the BGGD fits and the 2D isoprobability contours, the joint distributions of luminance subband responses are well modeled as bivariate generalized Gaussian. Moreover, Figure 5 also shows that there exist orientation-dependent dependencies between spatially adjacent subband responses since their correlations vary with the subband tuning orientations. For example, the correlation between horizontally adjacent subband responses is the strongest when the subband tuning orientation is equal to $\frac{1}{2}\pi$ (rad). In the next subsection, we present an NSS correlation model that captures these orientation dependencies.

Correlation NSS model

Here we model the correlations between the spatially neighboring, divisively normalized bandpass luminance responses described and demonstrated in the previous subsection. In particular, we have found that the correlation coefficients between spatially adjacent bandpass responses possess strong orientation dependencies (Su et al., 2014b, 2015a). For example, horizontally adjacent bandpass responses are most correlated when the subband tuning orientation aligns at $\frac{1}{2}\pi$ (rad), and become nearly uncorrelated at orientation 0 and π (rad). The correlation is periodic in the relative orientation between spatial and subband tuning orientation. This relative orientation regularity of correlation implies that there exist powerful constraints on spatially neighboring bandpass image responses.

Indeed, the periodic relative orientation dependency of the correlation coefficients between spatially adjacent bandpass responses can be well modeled in a closed form by an exponentiated sine function:

$$\rho = f(\theta_1, \theta_2) = A\left[\frac{1+\sin\left(\frac{2\pi(\theta_2-\theta_1)}{T}+\varphi\right)}{2}\right]^{\gamma} + c \tag{13}$$

where ρ is the correlation coefficient between spatially adjacent bandpass responses, θ_1 and θ_2 are the spatial and subband tuning orientations, respectively, A is amplitude, T is the period, φ is the phase, γ is an exponent, and c is the offset. We use the same definition of the spatial orientation θ_1 as in (Su et al., 2015a), where $\theta_1 = 0$ (rad) when the bandpass responses are sampled at vertically adjacent locations; for example, (x, y) and $(x, y+1)$, and $\theta_1 = \frac{1}{2}\pi$ (rad) when they are sampled at horizontally adjacent locations and, for example, (x, y) and $(x+1, y)$. When measured on naturalistic photographic images, the correlation coefficient is π-periodic, reaching maximum when $\theta_2 - \theta_1 = k\pi$, $k \in \mathbb{Z}$, yielding a three-parameter exponentiated cosine model:

$$\begin{aligned}\rho = f(\theta_1, \theta_2) &= A\left[\frac{1+\cos(2(\theta_2-\theta_1))}{2}\right]^{\gamma} + c \\ &= A[\cos(\theta_2-\theta_1)]^{2\gamma} + c\end{aligned} \tag{14}$$

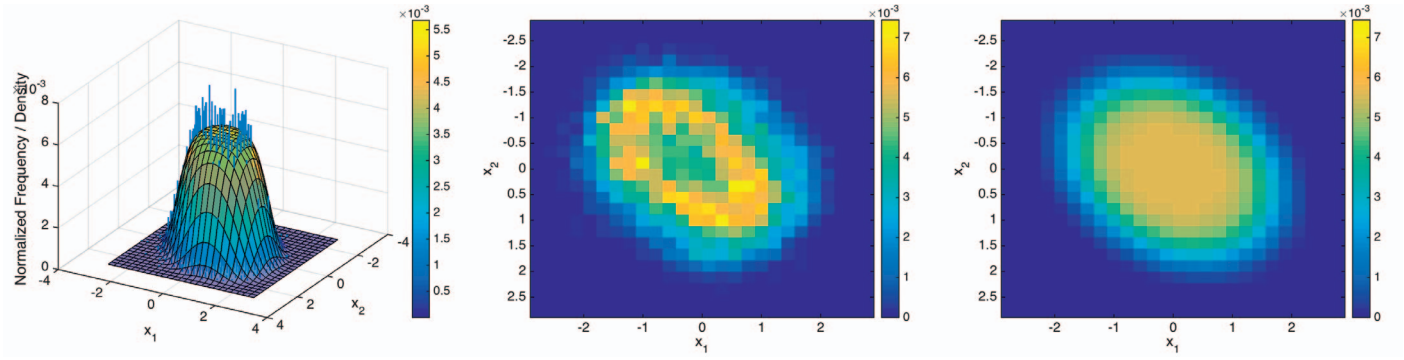
While the periodicity of the model is unsurprising, the specific and explicit parametric form of the model is unexpected. Indeed, it holds well for $\gamma = 1$ (Sinno & Bovik, 2015), although here we leave this parameter free, since it tends to vary with the quality of the picture. By computing the correlation coefficients between all horizontally adjacent subband responses within each image patch for all orientations at the same subband scale, and fitting each with the exponentiated cosine model, we arrive at three more parameters, A, γ, and c, that are descriptive of each patch's natural statistical structure, for each subband scale. All three NSS correlation model parameters from the two subband scales are included in our small feature set descriptive of each image patch. The fitting parameters are estimated via nonlinear least squares using the Levenberg-Marquardt algorithm (Marquardt, 1963).

Figure 6 plots an empirical correlation coefficient curve from the example image patch in Figure 3 as a function of $\theta_2 - \theta_1$ as well as its overlaid exponentiated cosine fit for horizontally adjacent subband responses; that is, $\theta_1 = \frac{1}{2}\pi$ (rad). The exponentiated cosine model nicely fits the spatial-oriented correlations between adjacent subband responses.

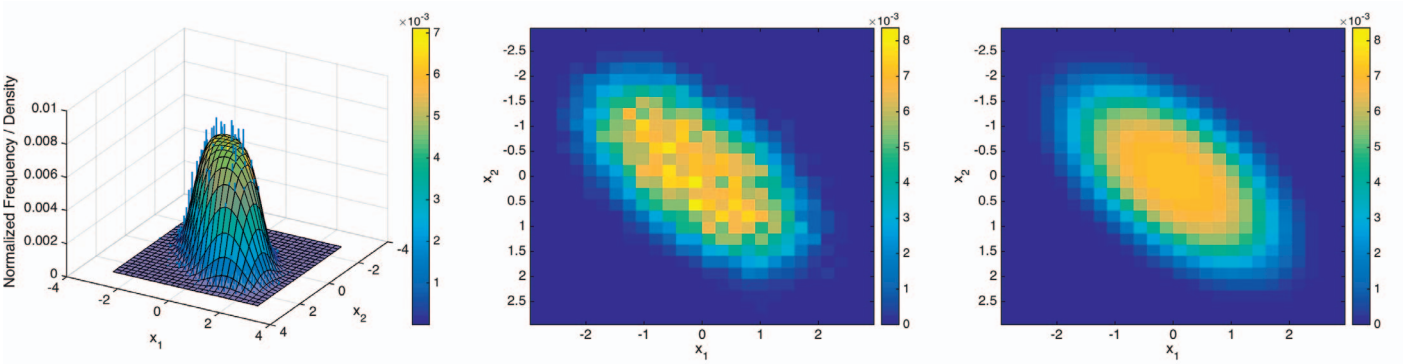
At this point, all of the NSS-based features that drive the proposed depth estimation model have been described. We thus define a 'depth-aware' image feature vector $\mathbf{f}_I$ to characterize each image patch:

$$\mathbf{f}_I = \left[\{\alpha_{u,s,r}, \beta_{u,s,r}\}, \{\alpha_{b,s,r}, \beta_{b,s,r}\}, \{A_s, \gamma_s, c_s\}\right]^{\top} \tag{15}$$

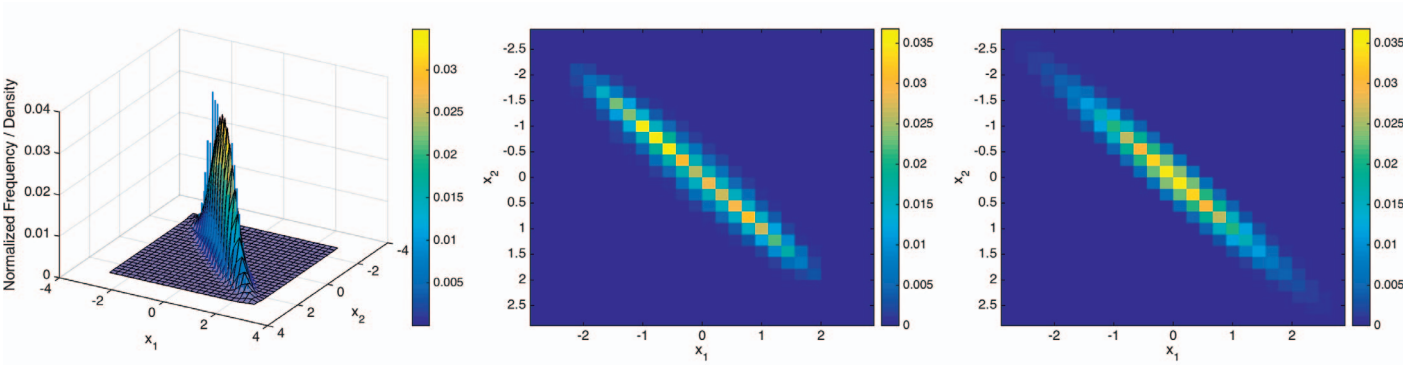
where $s \in \{1, 2, \ldots, S\}$, S is the number of scales, and $r \in \{1, 2, \ldots, R\}$, R is the number of subband orientations. In our model implementation, we use the subband responses from all eight subband orientations to extract the NSS image feature vector of each patch,



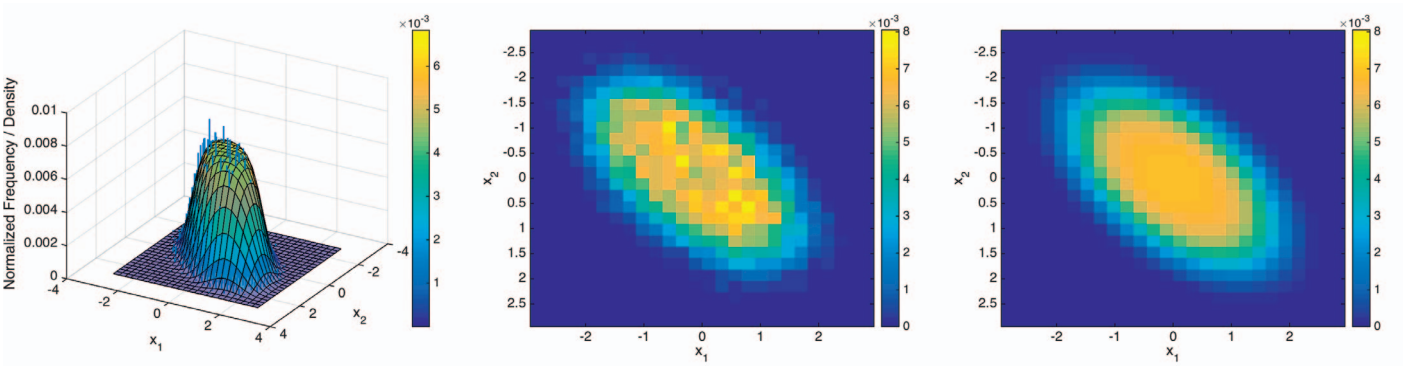
(a) 0 (rad) (b) 0 (rad) (c) 0 (rad)



(d) $\frac{1}{4}\pi$ (rad) (e) $\frac{1}{4}\pi$ (rad) (f) $\frac{1}{4}\pi$ (rad)



(g) $\frac{1}{2}\pi$ (rad) (h) $\frac{1}{2}\pi$ (rad) (i) $\frac{1}{2}\pi$ (rad)



(j) $\frac{3}{4}\pi$ (rad) (k) $\frac{3}{4}\pi$ (rad) (l) $\frac{3}{4}\pi$ (rad)

Figure 5. Joint histograms of horizontally adjacent responses from the example image patch in Figure 3 and the corresponding bivariate generalized Gaussian distribution (BGGD) fits for different subband tuning orientations. From top to bottom rows: 0 (rad), $\frac{1}{4}\pi$, $\frac{1}{2}\pi$, $\frac{3}{4}\pi$. The left column shows 3D perspective plots of the bivariate histograms and the corresponding best BGGD fits, where the blue bars are the empirical histogram values and the colored meshes represent the BGGD fits. The middle and right columns depict 2D isoprobability contour plots of the histograms and the BGGD fits, respectively.

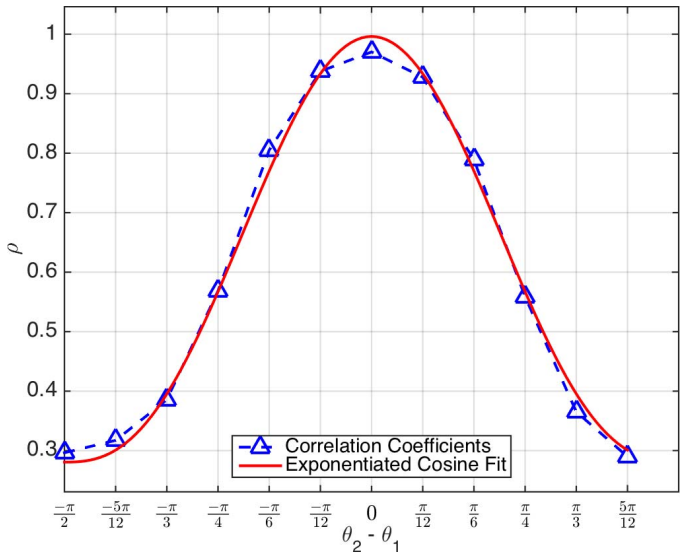


Figure 6. Graphs of the correlation coefficients between horizontally adjacent subband responses from the example image patch in Figure 3 and the corresponding best-fitting exponentiated cosine model. The correlation coefficient fits are plotted against $\theta_2 - \theta_1$ (rad), where $\theta_1 = \frac{1}{2}\pi$ (rad) is the horizontal tuning orientation, and $\theta_2 = 0, 1/12\pi, \ldots, 11/12\pi$ (rad) are the 12 subband tuning orientations.

resulting in a 38-dimensional feature vector $\mathbf{f}_I$, which is of length 2 (NSS univariate model parameters) $\times$ 8 (all eight subbands) + 2 (NSS bivariate model parameters) $\times$ 8 (all eight subbands) + 3 (NSS correlation model parameters) $\times$ 2 (two subband scales) = 38.

Depth feature extraction

In order to construct the priors and likelihoods of our Bayesian depth estimation model, as shown in Figure 1, it is also necessary to extract depth "features" from depth patches to characterize the depth structures constituting the prior, and to associate these with corresponding image patches to form the likelihood. To create this representation we perform a multiscale, multiorientation decomposition on the ground-truth depth patches, from which we extract depth features from the patterns of activation across different subbands. While there is no direct physiological evidence (that we are aware of) that depth maps are encoded in the primate cortex using something like a wavelet-based decomposition, it is not an unlikely supposition given the preponderance of bandpass processing of low-level visual data. This is even more likely given that depth information is being provided by bandpass V1 outputs (in our monocular depth estimation model). In this regard, there is evidence that cyclopean depth (i.e., the depth representation after disparity processing) is processed by orientation- and frequency-tuned bandpass channels, not unlike luminance information, but at a coarser spatial scale (Tyler, 1974, 1975, 1983; Schumer & Ganz, 1979; Cobo-Lewis & Yeh, 1994).

With this in mind, we compute divisively normalized steerable pyramid responses of the ground-truth depth patch as in Equation 7 to obtain an 8-bin histogram using the eight canonical orientation subbands; that is, $\theta = 0, \frac{1}{8}\pi, \ldots, \frac{7}{8}\pi$ (rad), at the first (finest) scale. We compute the average of all of the decomposed responses at each orientation subband to produce the corresponding histogram bin value.

In addition to the histograms of depth subband response magnitudes, we also compute the depth gradient histograms as part of our depth features. Specifically, we compute the local histograms of the depth gradient magnitude (Lowe, 1999) within eight orientation bins of debiased and normalized patch depths extracted from ground-truth depth maps. These histograms of bandpass, nonlinearly normalized depth gradient values are highly regular NSS signals that supply very fine-scale bandpass depth features. To obtain debiased and normalized depth patches, the patch mean value is subtracted from each depth patch and the result is then divisively normalized by the depth patch standard deviation.

The gradient vectors, $\nabla D = \left[g_{D_x} g_{D_y}\right]^\top = \left[\frac{\partial D}{\partial x} \frac{\partial D}{\partial y}\right]^\top = [(D(x+1, y) - D(x-1, y))\ (D(x, y+1) - D(x, y-1))]^\top$ at each coordinate of the resulting normalized depth patches D are computed (using the centered difference template $[-1\ 0\ 1]^\top$ along the two cardinal orientations) and projected onto each of the eight canonical orientations θ: $g_{D_x} \cos\theta + g_{D_y} \sin\theta$. The corresponding histogram is found for $\theta = 0, \frac{1}{8}\pi, \ldots, \frac{7}{8}\pi$ (rad), resulting in an 8-bin histogram for each depth patch, where each bin of the histogram represents the average of the depth gradient magnitudes projected onto the corresponding orientation over all pixel locations. In sum, a 16-dimensional depth feature vector $\mathbf{f}_D$ characterizing each depth patch is arrived at from the two 8-bin histograms described above: one of the gradient magnitudes computed from the debiased and normalized depth patch, and the other of the subband response magnitudes computed from the perceptually decomposed depth patch. These feature vectors are used to create prior and likelihood models, as explained in the next section.

Figure 7 shows the corresponding depth patch of the example image patch in Figure 3, including the ground-truth depth patch, the debiased and normalized depth patch, and the divisively normalized subband depth patches. Figure 8 plots the two 8-bin histograms computed from the depth patches shown in Figure 7. It can be seen that the two 8-bin histograms capture different aspects of the depth structure embedded in the patch.

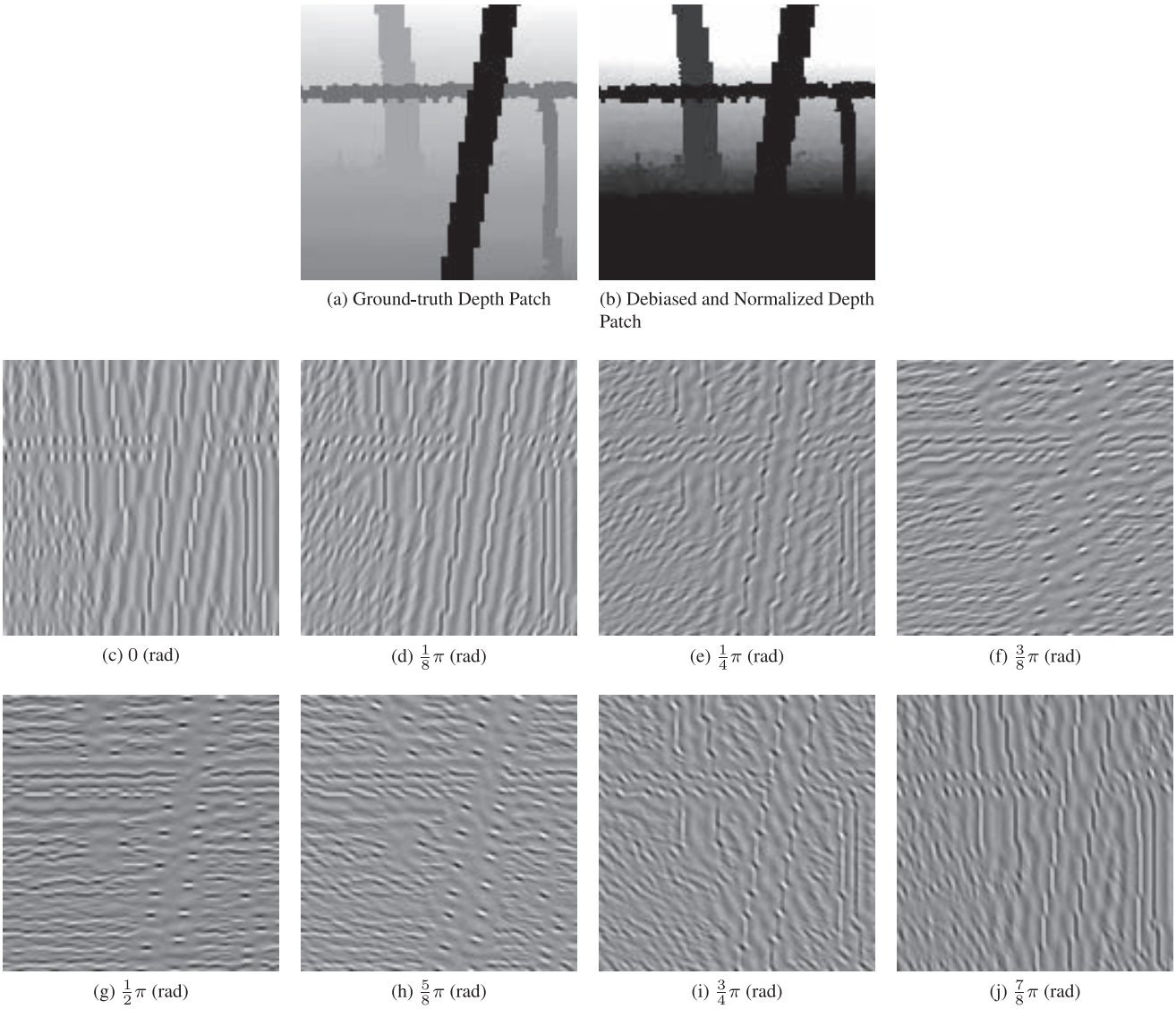
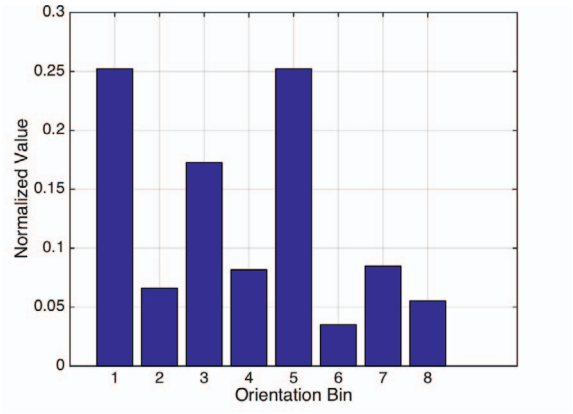


Figure 7. The depth patch corresponding to the example image patch in Figure 3. Top row: the ground-truth depth patch and the debiased and normalized depth patch. Middle and bottom rows: the perceptual decomposition (steerable pyramid subband responses after DNT) along the eight orientations used in the depth feature extraction process.

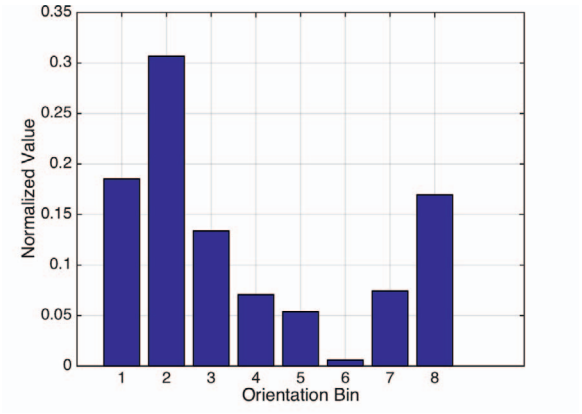
Before we explain the details of the prior and likelihood models, it will be beneficial to describe how the Bayesian inference (i.e., the estimated depth patch) is formed using the extracted NSS image features. As illustrated in Figure 1, the Bayesian model consists of priors and likelihoods, where the priors are a set of representative depth patterns/structures derived from the ground-truth depth patches, and the likelihoods are the conditional probability distributions of the extracted NSS image features given each prior. Therefore, the Bayesian model takes the extracted NSS feature of an image patch as input, computes the likelihood probability of that feature given all priors, combines the prior probability and the likelihood probability for each prior, and outputs the most probable depth patch, which is the representative depth pattern/structure of the prior having the highest posterior probability. Next, we explain how we derive the priors and likelihoods from the ground-truth depth patches and their associated image patches using the extracted NSS depth and image features.

Prior

It has been observed that discontinuities in depth maps are usually colocated with luminance edges occurring in the corresponding optical images (Jou & Bovik, 1989). Depth patches having similar depth patterns may be expected to exhibit similar luminance distributions (Y. Liu et al., 2011). In other words, some image patches may be distinctive enough that their



(a) 8-bin histogram from the debiased and normalized depth patch



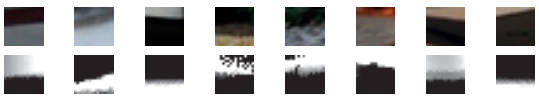
(b) 8-bin histogram from the divisively normalized sub-band depth patch

Figure 8. The two 8-bin histograms of depth gradient magnitudes, which form the 16-dimensional depth feature vector $\mathbf{f}_D$. The eight bins (from 1 to 8) in both histograms represent the eight subband orientations: 0, $\frac{1}{8}\pi, \ldots, \frac{7}{8}\pi$ (rad).

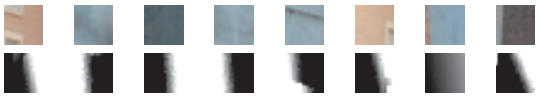
latent depth/3D structure can be predicted from their luminance appearance alone (Owens, Xiao, Torralba, & Freeman, 2013). Moreover, depth maps tend to possess simpler, more regular patterns than natural luminance images. Based on these observations, we built a dictionary of canonical depth patterns by clustering the processed, characteristic depth features ($\mathbf{f}_D$) extracted from the depth patches, as explained in the preceding section. As a simple method of data reduction, we employ the centroid-based k-means algorithm. The k-means algorithm is a simple clustering technique, whereby each sample of a set of n-dimensional data is assigned to one of k clusters according to a minimum Euclidean distance criterion (Lloyd, 1982).

Figure 9 shows examples of several canonical depth patterns (near cluster centroids) extracted by the k-means algorithm assuming five clusters, each with eight examples. For each canonical depth pattern, the bottom row shows the clustered depth patches (normalized residues) using the extracted features, while the top row shows the coregistered image patches. The depicted canonical depth patterns contain a variety of geometric structures, including depth discontinuities along the horizontal direction (pattern-1) and along the vertical direction (pattern-2); a smoother variation of depth along the horizontal direction (pattern-3) and along the vertical direction (pattern-4); and a busier, more complex pattern of depth changes (pattern-5). Complex depth patterns like pattern-5 are relatively

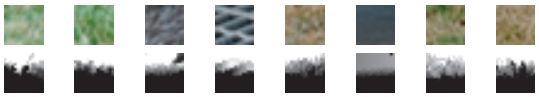
(a) Pattern-1



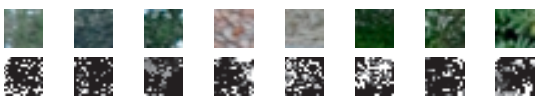
(b) Pattern-2



(c) Pattern-3



(d) Pattern-4



(e) Pattern-5

Figure 9. Examples of canonical depth patterns. For each canonical depth pattern, the bottom row shows the clustered depth patches (normalized residues) using the extracted features, while the top row shows the coregistered image patches. See text for more details.

uncommon, and appear in scenes containing rough objects, such as trees and grass. As the number of clusters is increased, these five canonical depth patterns still exist in similar form, although other clusters of depth patches emerge having similar structures that differ in some ways, such as orientation. In sum, the depth prior of the proposed Bayesian model consists of the normalized residual depth patch $\mathbf{d}_n$, (i.e., the cluster centroid associated with each canonical depth pattern), and the ratio $p(n)$ of each canonical depth pattern among all processed depth patches, where $n \in \{1, 2, \ldots, N\}$ and N is the number of canonical depth patterns (i.e., the number of clusters used by the k-means algorithm). Since increasing the number of clusters did not improve performance, in our implementation, we used $N = 5$ to keep the model as simple as possible. We study the effects of varying the number of canonical depth patterns in the discussion section.

The above procedure may be viewed as a way of finding a "sparse" set of representative depth patches. This suggests that a more sophisticated "sparse basis" might be found from which depth estimates could be computed. Here we used the k-means algorithm as a simple and efficacious proof of concept. The canonical depth patterns are conceptually similar to the idea of 3D primitives (Fouhey, Gupta, & Hebert, 2013). However, since our depth priors are constructed using NSS depth features obtained via a perceptual decomposition, as described in the Depth feature extraction subsection above, they are perceptually available, and can be acquired without solving regularized optimization problems.

Likelihood

As may be observed from the canonical depth patterns shown in Figure 9, depth discontinuities in range maps consistently align with luminance edges in coregistered natural images of the same scene (Jou & Bovik, 1989; Y. Liu et al., 2011). However, textured areas in photographic images that present significant variations in luminance/chrominance may not necessarily correspond to depth changes. In other words, high correlations exist between image edges and depth discontinuities, although the relationship is asymmetric. If the bandpass response to an image patch contains significant energy, then there is a relatively high likelihood of colocated variations (i.e., large depth gradients) in the corresponding range map. Conversely, if the range map contains large variations, then the colocated image bandpass response is even more likely to be large. To generalize and better utilize these relationships between image and depth variations in naturalistic settings, we derive a likelihood model that associates image patches with appropriate canonical depth patterns.

Assume that N canonical depth patterns have been obtained that define the prior using k-means clustering. Assign each image patch a label indicating its associated canonical depth pattern (cluster centroid) for its corresponding depth patch. Then, using these labeling results, the depth-aware feature vectors, i.e., $\mathbf{f}_I$ in Equation 14, that are extracted from each image patch are used to train a classifier using a multivariate Gaussian mixture (MGM) model. The reason that the MGM model is well suited to this classification task is that, as may be observed in Figure 9, image patches presenting different appearances and/or textured surfaces may yet be associated with the same canonical depth pattern. Therefore, we exploit a simple multimodal Gaussian mixture model trained on each canonical depth pattern to handle the heterogeneity of its image patches. As a result, the number of mixtures used to train the MGM classifier is the same as the number of clusters used in the k-means algorithm for learning the canonical depth patterns. An MGM model is defined as:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (16)$$

where is the model parameter vector, $\mathbf{x}$ is a multidimensional data vector (e.g., some measurement or a feature), $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the m-th Gaussian component, and w_m is the m-th mixture weight with the constraint that $\sum_{m=1}^{M} w_m = 1$. Note that the complete MGM model is parameterized by $\theta = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, m \in \{1, \ldots, M\}$, which includes the mean vectors, covariance matrices, and mixture weights from all Gaussian components. Finally, the m-th Gaussian component density function is given by:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{K/2} |\boldsymbol{\Sigma}_m|^{1/2}} e^{\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^\top \boldsymbol{\Sigma}_m{}^{-1} (\mathbf{x}-\boldsymbol{\mu}_m)\right]} \quad (17)$$

where K is the dimensionality of $\mathbf{x}$. Here the depth-aware feature vector is modeled: $\mathbf{x} = \mathbf{f}_I \in \mathbb{R}^K$. For each canonical depth pattern, an MGM model is created using the feature vectors extracted from all of the image patches within the pattern. Therefore, the likelihood of encountering an image patch with a specific extracted feature $\mathbf{f}_I$ given a particular canonical depth pattern indexed by n can be expressed as:

$$p(\mathbf{f}_I; \boldsymbol{\theta}_n) = \sum_{m=1}^{M} w_{n,m} \mathcal{N}\left(\mathbf{f}_I; \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m}\right) \quad (18)$$

where $\boldsymbol{\theta}_n = \left\{w_{n,m}, \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m}\right\}, m \in \{1, \ldots, M\}$, and

$n \in \{1, \ldots, N\}$. In our implementation, we set M equal to the number of canonical depth patterns (i.e., $M = N = 5$) to handle the heterogeneity of the associated image patches in each depth prior, and to be able to estimate the MGM model parameters, $\boldsymbol{\theta_n}, n \in \{1, \ldots, N\}$, using an iterative expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

Regression on mean depth

As discussed in the Depth feature extraction section, a preprocessing step is performed prior to the extraction of features from depth patches to learn the prior, whereby each depth patch is normalized by removing the mean and standard deviation to homogenize the depth patterns, and to better reveal their essentially distinguishing characteristics. In order to be able to add the mean value of each depth patch back when estimating the true range values of test image patches, it is necessary to learn a mapping from the image feature space using a regression model. In other words, given an input image patch, the trained regressor can be used to estimate the mean range of the corresponding depth patch using the extracted depth-aware image feature vector $\mathbf{f}_I$. Since we observed negligible influences, both numerically and visually, of patch standard deviations on the estimated depth maps, the proposed Bayesian model is able to attain the same degree of performance without recovering depth patch standard deviations.

In addition to $\mathbf{f}_I$, we exploit two other useful monocular depth cues to assist with recovery of true range values. The experiments conducted in Schwartz and Sperling (1983) and Dosher, Sperling, and Wurst (1986) suggested that the positive proximity luminance covariance (PLC; i.e., the observation that objects closer to the observer are made brighter), may serve as an important cue in the perception and interpretation of 3D structures by human subjects. Potetz and Lee (2003) showed that there exists a general dependency between intrinsic image brightness and colocated distance in natural scenes. We use this "the brighter the nearer" law to further guide the estimation of the mean patch depth value using the average luminance of the corresponding image patch. Moreover, in natural scenes, the distance from the nodal point to any point in the scene tends to increase as its height increases. Specifically, if we assume that the y-coordinate of a pixel increases from the bottom to the top of an image, the range values of pixels with larger y-coordinates are generally larger than those with smaller y-coordinates. Thus, we introduce as a second additional feature into the regressor on mean depth values, the normalized y-coordinate of each patch in the image:

$$f_y = \frac{p_y}{I_h} \quad (19)$$

where p_y is the y-coordinate of the image patch, and I_h is the height of the image. Thus, in sum, the aggregate feature vector characterizing each image patch used in the regression model to learn mean patch depth values includes the depth-aware feature set $\mathbf{f}_I$, the average patch luminance, and the normalized y-coordinate, f_y. In the proposed Bayesian model, we utilize a standard support vector regressor (SVR; Schölkopf, Smola, Williamson, & Bartlett, 2000) to implement the training and testing processes, using multiple train–test sets as described in the section of experimental results. SVR is generally noted for being able to effectively handle high dimensional data (Burges, 1998). We implemented the SVR model with a radial basis function (RBF) kernel using the LIBSVM package (Chang & Lin, 2011).

Bayesian model

We now describe how the proposed Bayesian framework incorporates the canonical depth pattern prior model, the likelihood model that associates image patches with different canonical depth patterns, and the regression model that recovers mean patch depth values. Given an input image, the model algorithm first divides it into overlapped patches of size $P \times P$, where a $\frac{1}{4}$ overlap (stride) is used (i.e., the patches overlap each other by $\frac{P}{4}$ pixels along both dimensions). In our implementation, we chose $P = 32$, although we have found the model to be robust to this choice, as we show later. We have also performed a thorough analysis with detailed discussion on the effect of the patch size, which will be covered in the discussion section. Next, the depth-aware feature vector $\mathbf{f}_I$ is extracted from each image patch, as well as the average luminance and the normalized y-coordinate, which are used, as described earlier, for mean depth regression. Then, the extracted feature vector $\mathbf{f}_I$ is fed into the trained prior, likelihood, and regression models to form a Bayesian inference of the corresponding estimated depth patch. Specifically, the estimated depth patch $\mathbf{D}$ of an image patch is formed as follows:

$$\mathbf{D} = \mathbf{d}_n + \mu_n \quad (20)$$

where $\mathbf{d}_n$ (obtained in the prior model) is the normalized residual depth patch associated with the estimated canonical depth pattern n, μ_n is the corresponding mean depth value obtained from the regression model, and n is the index of the estimated canonical depth pattern derived from the prior and likelihood models, which is given by:

$$n = \underset{n'}{\operatorname{argmax}}\{p(n'|\mathbf{f}_I)\} = \underset{n'}{\operatorname{argmax}}\{p(\mathbf{f}_I|n')p(n')\} = \underset{n'}{\operatorname{argmax}}\{p(\mathbf{f}_I;\boldsymbol{\theta}_{n'})p(n')\} \quad (21)$$

where $p(\mathbf{f}_I|n') = p(\mathbf{f}_I;\boldsymbol{\theta}_{n'})$ is the likelihood (Equation 18) of encountering an image patch having the extracted feature vector $\mathbf{f}_I$ given a canonical depth pattern n', and $p(n')$ is the corresponding prior probability (ratio) of the estimated canonical depth pattern n'.

Stitching

The last stage of the overall depth estimation system is to stitch all of the depth patches together to create a final estimated depth map using only the monocular test image as input. Seeking simplicity, we define the stitching operation to simply average the estimated depth values of overlapped pixels across the assembled depth patches.

Experimental results

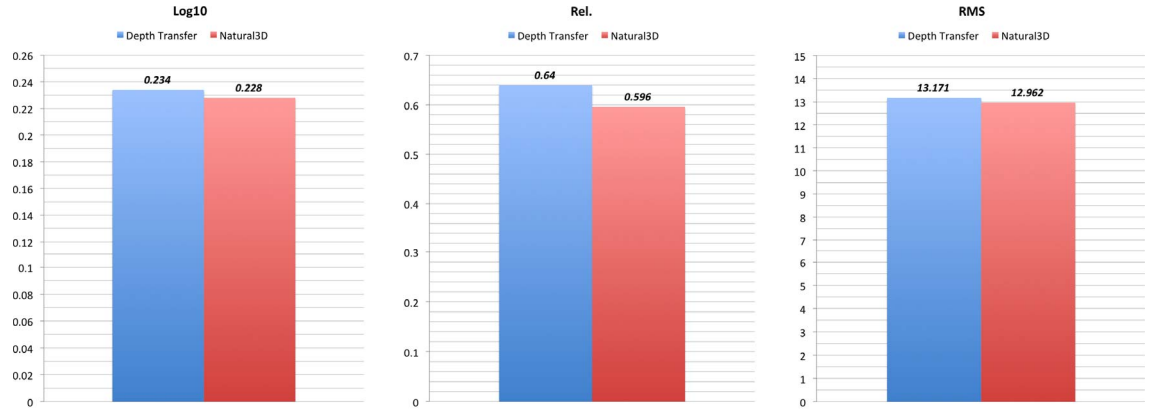
To evaluate its performance in estimating depth from single monocular images, we trained and tested the proposed Bayesian model extensively on three publicly accessible databases, the LIVE Color+3D Database Release-2 (Su et al., 2016b), the Make3D Laser+Image Dataset-1 (Saxena, Chung et al., 2005; Saxena, Sun et al., 2005; Saxena et al., 2009), and the NYU Depth Dataset V2 (Silberman, Hoiem, Kohli, & Fergus, 2012; Silberman, Kohli et al., 2012). For the sake of brevity, we hereafter call our proposed Bayesian model Natural3D.

Databases

The LIVE Color+3D Database Release-2 consists of 99 pairs of color images and accurately coregistered ground-truth depth maps, all with a high-definition resolution of 1920×1080. We constructed the database using an advanced range scanner, RIEGL VZ-400, with a 12.1 megapixel Nikon D700 digital single-lens reflex camera mounted on top of it (RIEGL Laser Measurement Systems, 2009). The RIEGL VZ-400 allows for a maximum scan angle range of 100° ($+60^\circ/-40^\circ$) and 60° in the vertical and horizontal direction, respectively, with a minimum angle step-width of 0.0024°. Scan speeds up to 120 lines/s can be achieved, with an angle measurement resolution of better than 0.0005° and a maximum measurement range of up to 500 m. Careful and precise calibration was executed before data acquisition, and a perspective transformation was applied to project the 3D point clouds of depth measurements onto the 2D image plane. We also took into account the real-lens distortions (viz., radial and tangential) to achieve accurate depth-image registration (Intel Corporation, 2000). The dense, ground-truth precision depth data we acquired in this way make the LIVE Color+3D Database Release-2 science-quality. As a result, the LIVE Color+3D Database Release-2 provides a rich source of information regarding natural depth statistics, and also serves as an excellent resource for evaluating depth estimation algorithms (including binocular, since coregistered stereo pairs are included). To avoid overlap between training and testing image/depth content, we split the entire database into 80% training and 20% testing subsets at each train–test iteration with no content shared between the training and testing subsets. This train–test procedure was repeated 50 times to ensure that there was no bias introduced due to image/depth training content.

The Make3D Laser+Image Dataset-1 contains a total of 534 pairs of color images and corresponding ground-truth depth maps, where 400 are used for training and 134 for testing with no content overlap. The color images are high resolution 2272×1704, while the ground-truth depth maps are only available at a very low resolution of 55×305. These very low resolution, sparse, ground-truth depth maps with unmatched aspect ratio to the color images make the Make3D Laser+Image less than ideal for developing and testing contemporary dense depth estimation algorithms. However, due to its early availability, it has been widely used for evaluating monocular depth estimation methods. To make a complete comparison, we also trained and tested Natural3D on the Make3D database.

The NYU Depth Dataset V2 is comprised of 1,449 pairs of aligned color images and dense depth maps taken from a variety of real-world indoor scenes. The database was recorded using the RGB and depth cameras of the Microsoft Kinect device (Microsoft, 2010). Both the color images and depth maps are available at the VGA resolution of 640×480, while the depth values fall within the range of 0.7–6.0 m due to the sensor limit. Kinect depth data is of low resolution, is notoriously noisy, and is limited to indoor scene acquisition. Nevertheless, even though the data are not science quality, it has been widely used in the computer vision literature for algorithm training and comparison; hence, we also test our model on these data. We used the raw depth maps to avoid any unnatural filling or inpainting of missing depth values from the sensor. As was done on the LIVE Color+3D Database Release-2, we performed 50 random train–test splits with 80% training and 20% testing on the entire dataset to avoid content bias in our experiments.



(a) Log10 (b) Rel. (c) RMS

Figure 10. The average metric performance comparison on the LIVE Color+3D Database Release-2.

Performance comparison

We compared Natural3D with a top-performing state-of-the-art depth estimation method, called Depth Transfer (Karsch et al., 2012) on all three datasets. Depth Transfer has delivered the best-reported performance on the Make3D Laser+Image Dataset-1. Depth Transfer first selects candidates from a database by matching a high-level image feature, GIST (Oliva & Torralba, 2001), and then optimizes an energy function to generate the most likely depth map by considering all of the warped candidate depth maps under a set of spatial regularization constraints. We also report results from two other recent computer vision methods on the very large NYU dataset. Im2Depth (Baig et al., 2014) uses a sophisticated sparse dictionary approach to transform RGB to depth, while in Eigen et al. (2014), a multiscale deep learning network using two deep network stacks is employed. Neither method uses perceptually relevant features beyond RGB pixel values (viz., the deep learner finds its own features). Unfortunately, these models are not reproducible on the high-quality LIVE Color+3D Database, which contains fine, detailed, naturalistic outdoor scene data rather than very low-resolution data (Make3D) or smooth indoor data (smooth floors, walls, and furniture in NYU). They are not available for testing and, in any case, training a deep learner would require vastly more data than is available in any outdoor/high-resolution 3D dataset.

Quantitative evaluation

We performed a quantitative evaluation on the two examined monocular depth estimation algorithms using three different objective metrics. We first report the results obtained using three common error metrics, the log error (Log10; B. Liu et al., 2010; Karsch et al., 2012; Baig et al., 2014):

$$\sum_{i=1}^{S} \frac{|\log_{10}\mathbf{D}(x_i,y_i) - \log_{10}\mathbf{D}^*(x_i,y_i)|}{S} \quad (22)$$

the relative error (Rel.; B. Liu et al., 2010; Karsch et al., 2012; Baig et al., 2014):

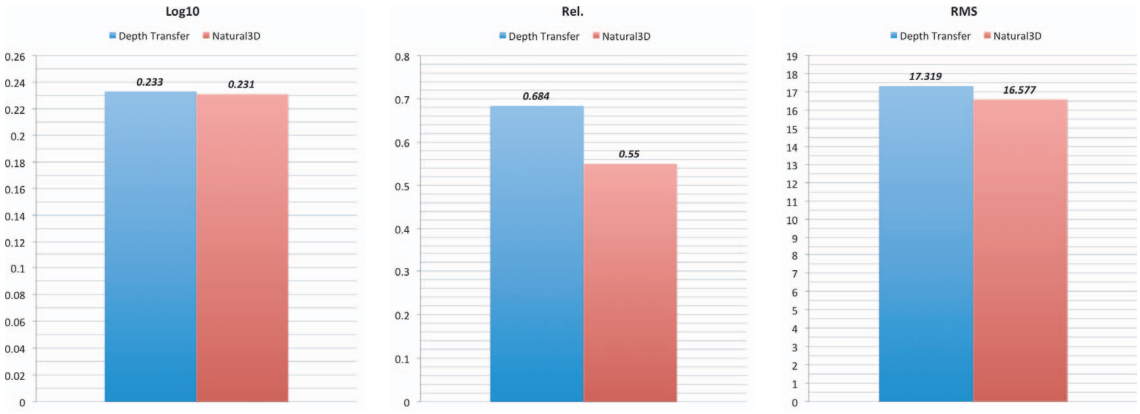
$$\sum_{i=1}^{S} \frac{|\mathbf{D}(x_i,y_i) - \mathbf{D}^*(x_i,y_i)|/\mathbf{D}^*(x_i,y_i)}{S} \quad (23)$$

and the root-mean-square error (RMS; Saxena et al., 2009; Fouhey et al., 2013; Baig et al., 2014):

$$\sqrt{\sum_{i=1}^{S} \frac{[\mathbf{D}(x_i,y_i) - \mathbf{D}^*(x_i,y_i)]^2}{S}} \quad (24)$$

where $\mathbf{D}(x_i,y_i)$ and $\mathbf{D}^*(x_i,y_i)$ are the estimated and ground-truth depth map at pixel location (x_i,y_i), respectively, and S is the number of pixels. Note that the ground-truth depth maps in the tested databases are measured in units of meters (m), as are the error metrics.

Figures 10 through 12 show the average metric performances on the LIVE Color+3D Database Release-2 (across all images in the train–test splits), the Make3D Laser+Image Dataset-1 (across all test scenes), and the NYU Depth Dataset V2 (across all images in the train–test splits), respectively, including those published results for Im2Depth (Baig et al., 2014) and Eigen et al. (2014). Natural3D achieves superior or similar performances to Depth Transfer in terms of all three error metrics, and is highly competitive with the heavily optimized and data-dependent computer vision methods, Im2Depth and Eigen et al. (2014), despite its use of simple, perceptually relevant NSS features. Figures 13 to 15 show the standard deviations of the three error metrics on all 3D databases for Natural3D



(a) Log10 (b) Rel. (c) RMS

Figure 11. The average metric performance comparison on the Make3D Laser+Image Dataset-1.

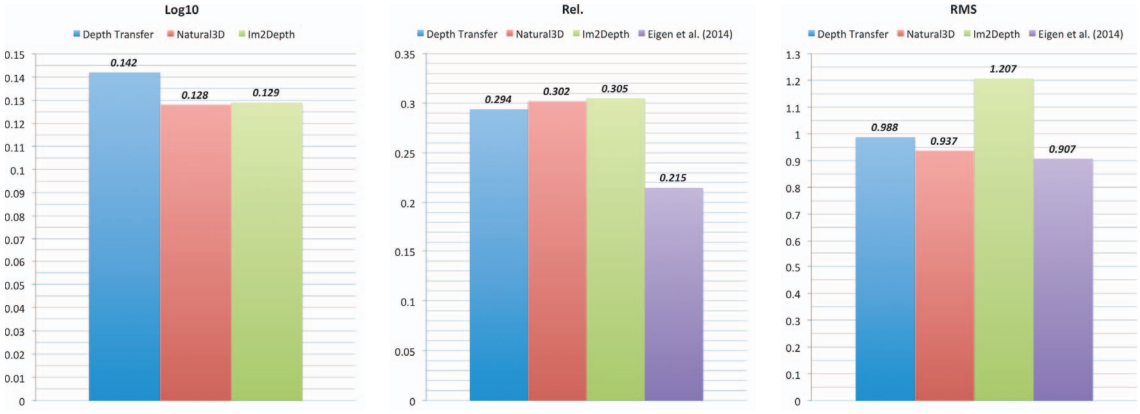
and Depth Transfer,[3] which reflects the performance consistencies of the examined depth estimation algorithms. Natural3D delivers more consistent performance in terms of Log10, while providing similar or better Rel. and RMS performances than Depth Transfer.

Visual examination

In addition to the quantitative comparison, we also supply a visual comparison by showing examples of the depth maps estimated by Natural3D and Depth Transfer along with the corresponding ground-truth depth maps, as shown in Figures 16 and 17 (from the Make3D Laser+Image Dataset-1), Figures 18 and 19 (from the LIVE Color+3D Database Release-2), and Figures 20 and 21 (from the NYU Depth Dataset V2). Note that for the Make3D Laser+Image Dataset-1, the ground-truth and estimated depth maps are scaled to match the image resolution for display purposes. We also supply scatter plots between the estimated and the ground-truth range values to gain a broader perspective of performance.

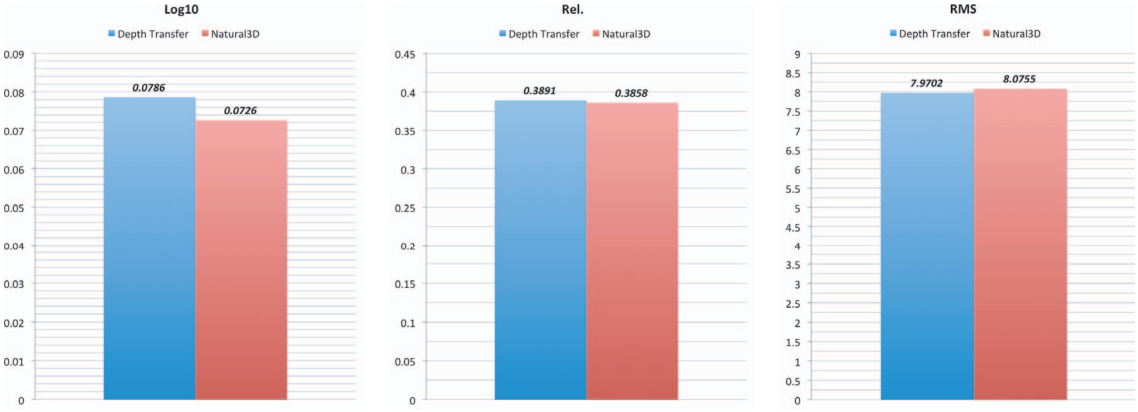
As may be seen on all three databases, Depth Transfer tends to over-smooth the estimated depth maps due to its smoothness constraint, while Natural3D is able to capture detailed depth structures in the scene. For example, in Figure 16, Depth Transfer incorrectly wipes out the house on the left, while Natural3D is able to separate the ground and the building. Similarly, the trees and sky in the background of Figure 16 are missing in the estimated range map delivered by Depth Transfer, while Natural3D successfully reconstructs most of them. In Figure 17, Natural3D is capable of recovering the tree depth structures, as well as identifying the sky in the background, while Depth Transfer only captures the ground.

As shown in Figure 18, Depth Transfer is not able to capture the tree trunks in the foreground, and it also incorrectly merges the tree trunks in the background. By comparison, Natural3D creates a



(a) Log10 (b) Rel. (c) RMS

Figure 12. The average metric performance comparison on the NYU Depth Dataset V2. The Log10 error metric is not reported in Eigen et al. (2014).



(a) Log10 (b) Rel. (c) RMS

Figure 13. The standard deviation of error metrics on the LIVE Color+3D Database Release-2.

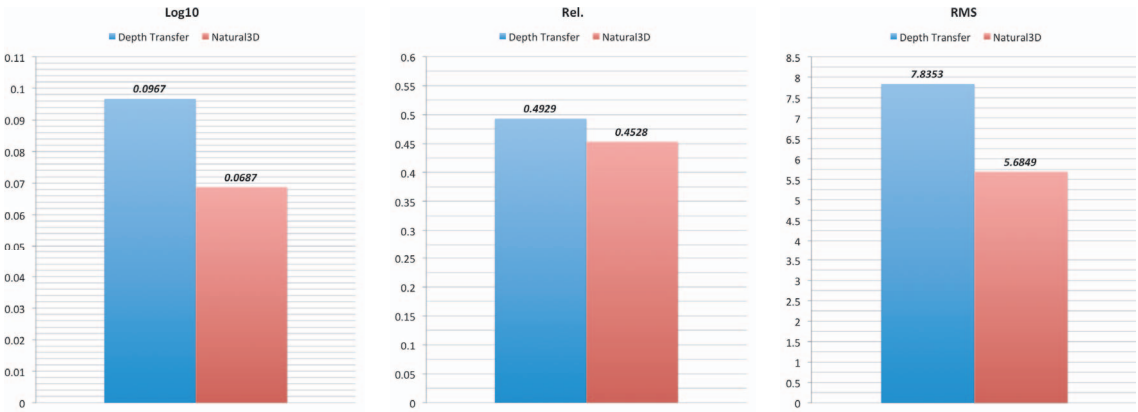
clearer representation of the foreground tree trunks. Figure 19 shows a number of human objects and a tree branch, posing more challenging content for monocular depth estimation algorithms. Natural3D successfully captures details such as the intersection of the human hand and the tree branch, while Depth Transfer fails to recover such complicated structures due to over-smoothing.

Figures 20 and 21 show example results of the estimated depth maps from the NYU Depth Dataset V2. Note that since indoor scenes are generally less textured than outdoor natural scenes, it is generally more difficult for algorithms to make accurate depth estimates without incorporating special features (such as planar fits) or heavily training on that type of data. However, as shown in Figure 21, Natural3D is still able to infer the relative depths of the objects on the table on the left side of the image, while Depth Transfer incorrectly smoothes them with the background.

It is remarkable that the simple, perceptually driven NSS-based approach Natural3D is able to compete so well on the indoor-only NYU dataset, given that the competitive models either deploy a smoothness constraint as in Depth Transfer, a dictionary constructed from indoor depth data (Im2Depth), or deep machine learning on rich indoor depth data.

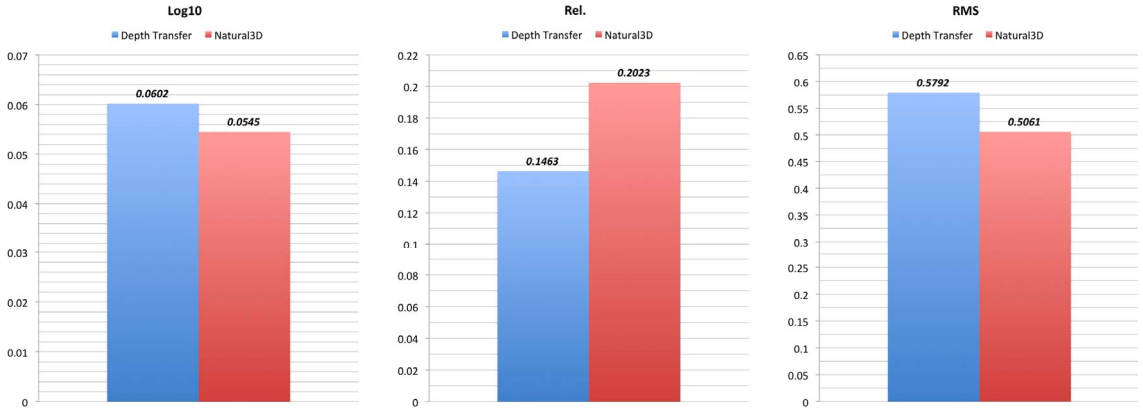
The complete experimental results of the two, including quantitative and visual comparison, examined monocular depth estimation algorithms on every image in both databases can be found at (Su, Cormack, & Bovik, 2016a).

Another advantage of Natural3D is that there is no need for an iterative solution process, resulting in greatly reduced computational complexity. Table 1 shows the runtime per estimated depth map for two examined algorithms. Natural3D and Depth Transfer were implemented using the MATLAB programming language, and the simulations were run on an Intel Core i7 quad-core processor with 16 GB memory. Since Natural3D utilizes trained prior and likelihood models, it runs almost 10 times faster than Depth Transfer, which uses an iterative procedure to solve an optimization function.



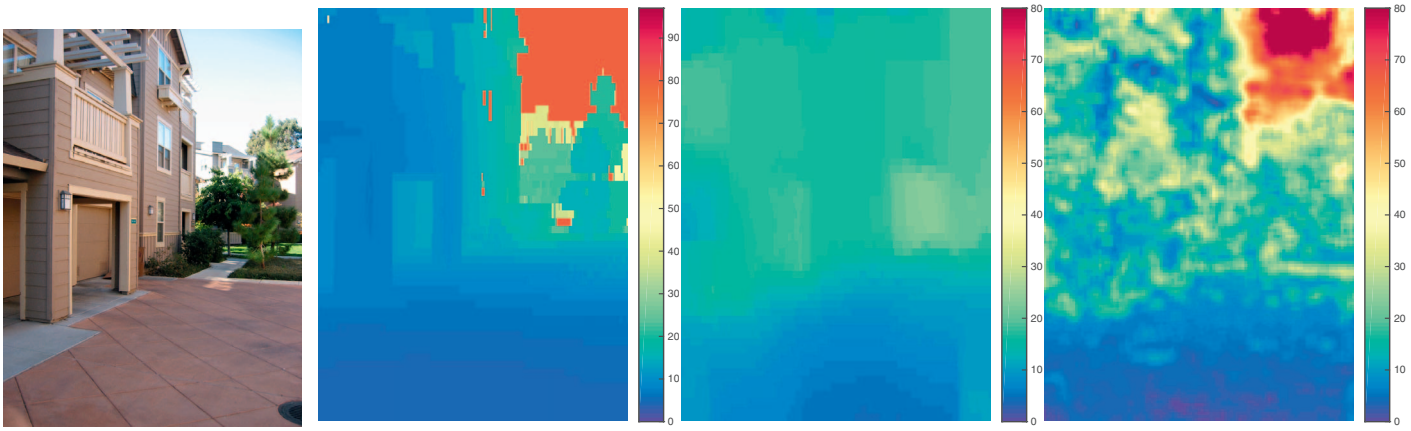
(a) Log10 (b) Rel. (c) RMS

Figure 14. The standard deviation of error metrics on the Make3D Laser+Image Dataset-1.

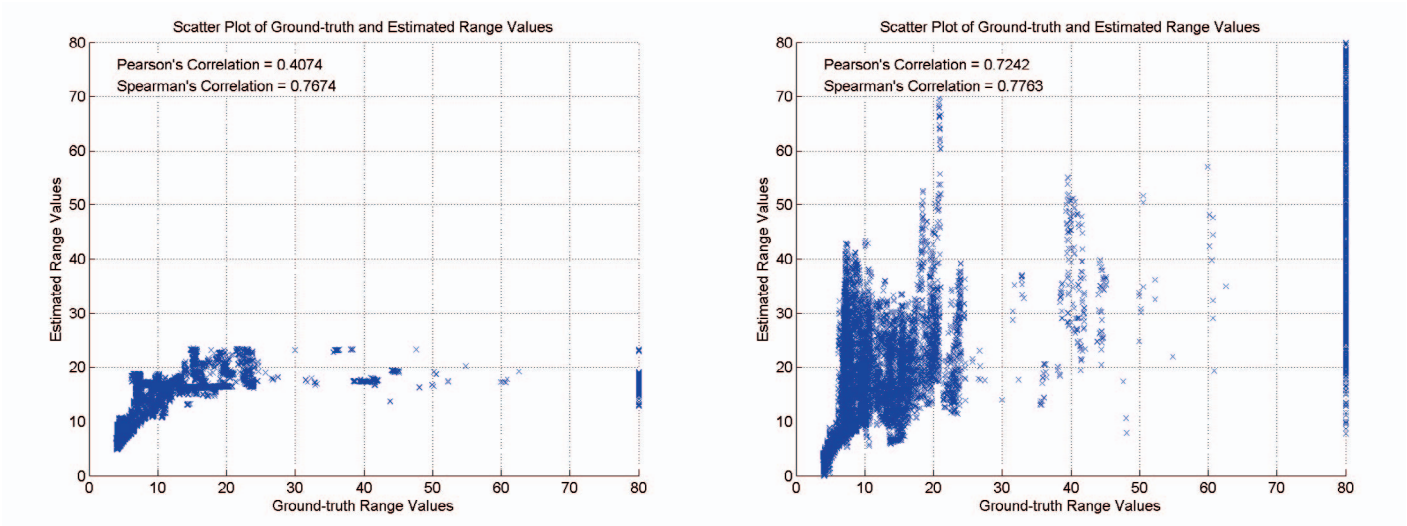


(a) Log10 (b) Rel. (c) RMS

Figure 15. The standard deviation of error metrics on the NYU Depth Dataset V2.

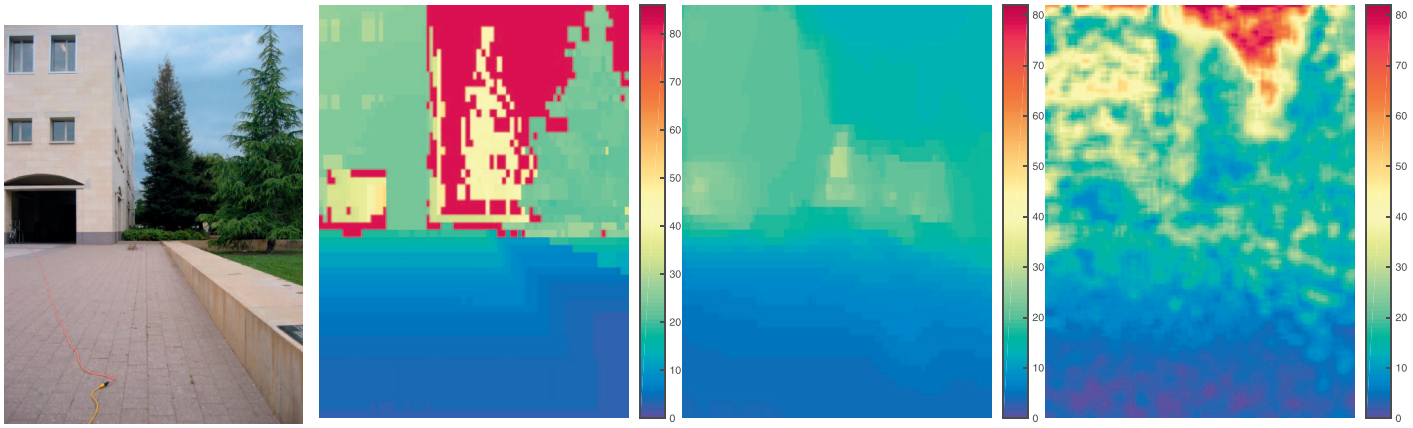


(a) Natural image (b) Ground-truth depth map (c) Estimated depth map by Depth Transfer (d) Estimated depth map by Natural3D

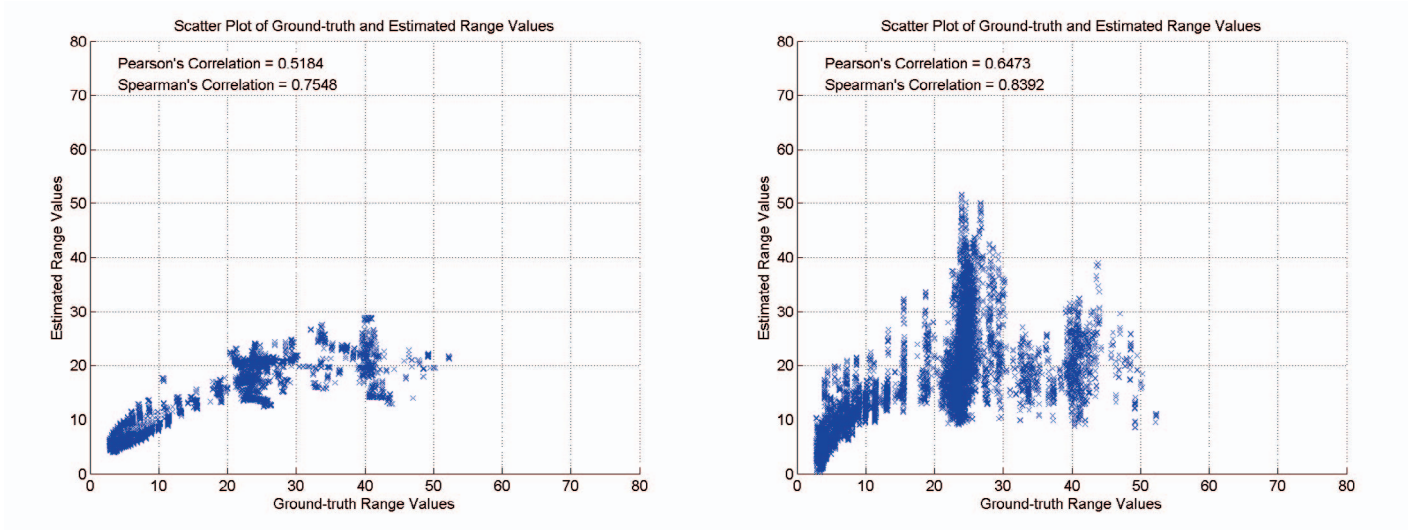


(e) Scatter plot of Depth Transfer result (f) Scatter plot of Natural3D result

Figure 16. Example estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.



(a) Natural image (b) Ground-truth depth map (c) Estimated depth map by Depth Transfer (d) Estimated depth map by Natural3D



(e) Scatter plot of Depth Transfer result (f) Scatter plot of Natural3D result

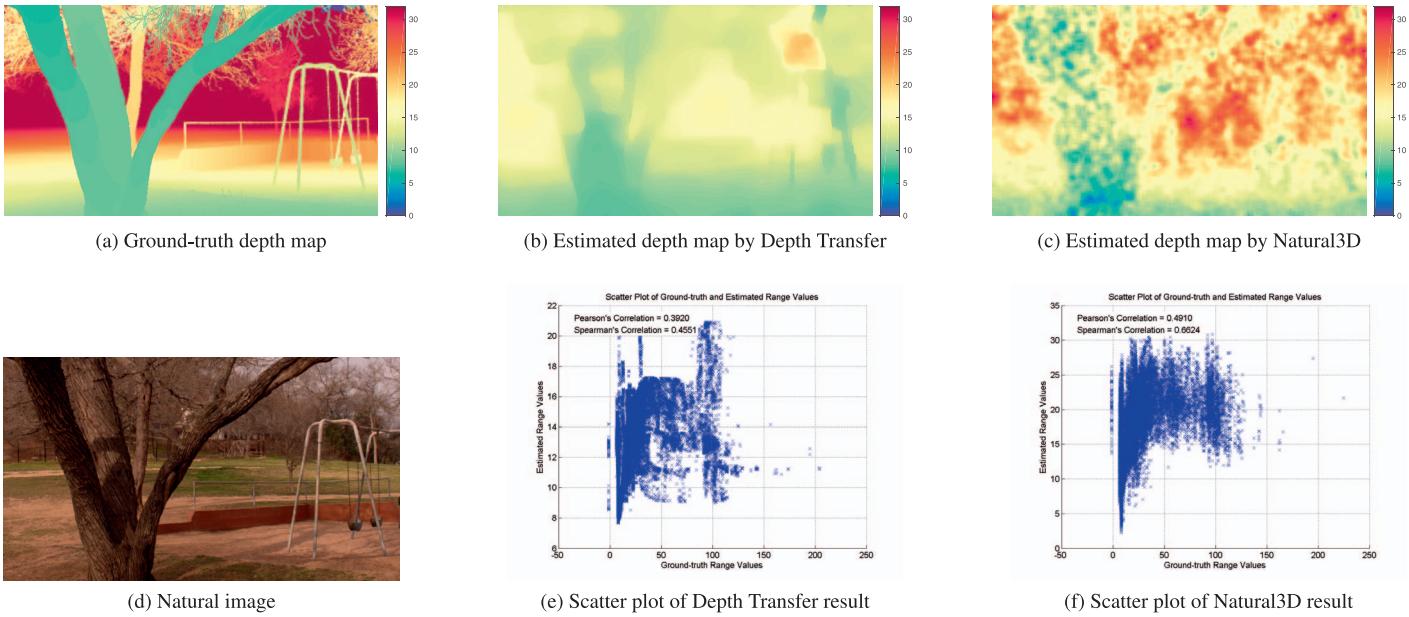
Figure 17. Example estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.

Discussion

In order to acquire a better understanding of the contributions of the individual components of Natural3D, as well as the power of the bivariate and correlation NSS models for depth estimation, we performed a thorough set of intrinsic analyses of the different algorithmic aspects of our NSS-based Bayesian framework. We first examine the choices of the two parameters in Natural3D, the patch size and the number of canonical depth patterns. We then evaluate variants of the framework to analyze the importance of each component, including the regression model, the NSS feature, and the Bayesian inference. Note that we performed the following intrinsic analyses on the LIVE Color+3D Database Release-2 because of the science-quality, high-resolution color images, and high-quality depth maps that it provides.

Patch size

In our implementation of Natural3D, we chose to use patches of size 32×32 ($P = 32$). This choice was based on the need to acquire enough bandpass response samples to obtain a reliable two-dimensional histogram, upon which accurate NSS models can be built. To support our choice of patch size, we trained and tested Natural3D using a variety of different patch sizes, and plotted the results in Figure 22 with three error metrics as a function of patch size. It may be seen that both the RMS and Rel. error metrics improved as the patch size was increased up to 32, while there was little variation in the computed Log10 error metric values. Moreover, both the RMS and Rel. error metrics stabilized for patch sizes larger than 32. This result indicates that once there are enough samples to construct reliable bivariate and correlation NSS models, the extracted features become sufficiently stable to enable Natural3D to deliver accurate and consistent



(a) Ground-truth depth map (b) Estimated depth map by Depth Transfer (c) Estimated depth map by Natural3D

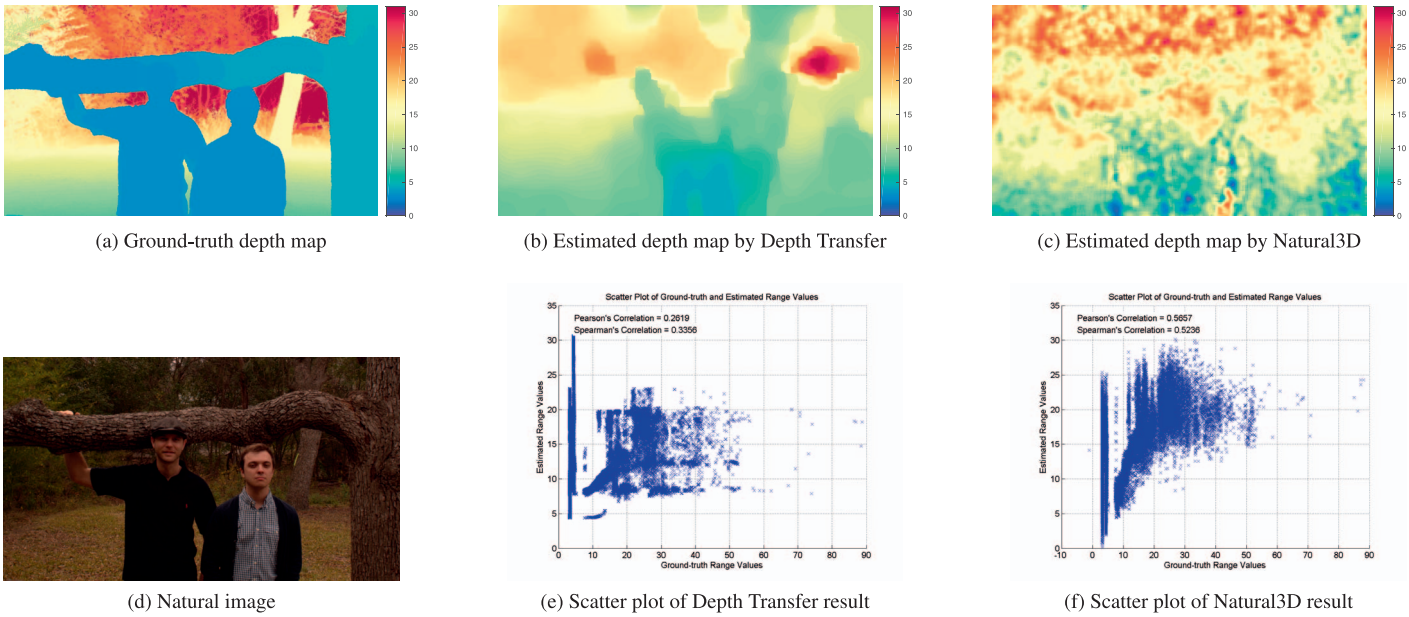
(d) Natural image (e) Scatter plot of Depth Transfer result (f) Scatter plot of Natural3D result

Figure 18. Example estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2.

depth estimation performance. In other words, the proposed framework is robust to the choice of patch size larger than 32.

Number of canonical depth patterns

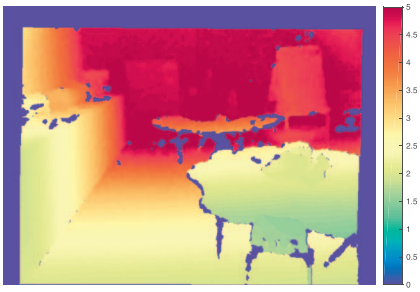
Another parameter choice we made in our implementation of Natural3D is the number of canonical depth patterns, which are the number of clusters used in the centroid-based k-means algorithm for learning the depth prior. To demonstrate the influence of the number of canonical depth patterns on the performance of Natural3D, we trained and tested the algorithm using different numbers of clusters in the k-means algorithm, and plotted in Figure 23 the three error metrics as a function of the number of canonical depth patterns. It can be seen that, while the relative



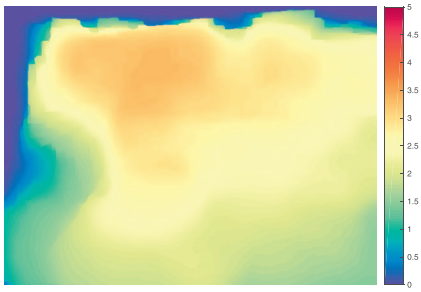
(a) Ground-truth depth map (b) Estimated depth map by Depth Transfer (c) Estimated depth map by Natural3D

(d) Natural image (e) Scatter plot of Depth Transfer result (f) Scatter plot of Natural3D result

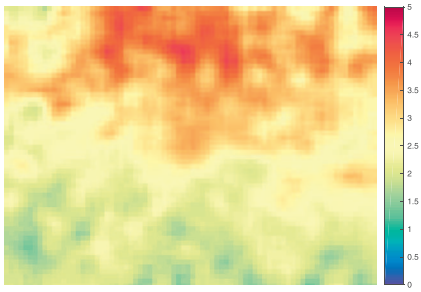
Figure 19. Example estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2. Photo credit: Dr. Brian McCann; pictured are Dr. Johannes Burge and Dr. Steve Sebastian. Dr. McCann, Dr. Burge, and Dr. Sebastian all are creators of LIVE Color+3D Database Release-2 and were members of Center for Perceptual Systems at The University of Texas at Austin.



(a) Ground-truth depth map

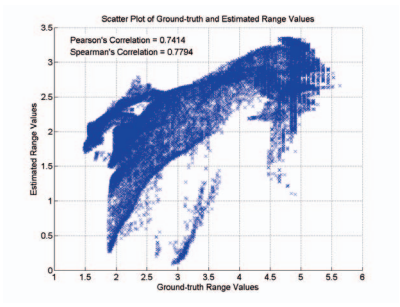


(b) Estimated depth map by Depth Transfer

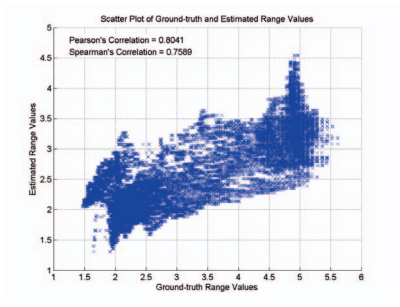


(c) Estimated depth map by Natural3D

(d) Natural image



(e) Scatter plot of Depth Transfer result

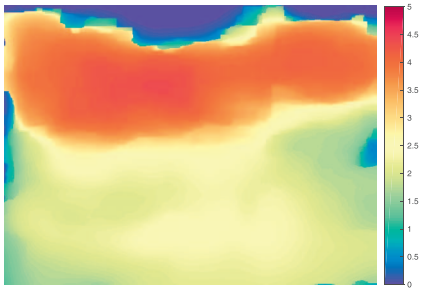


(f) Scatter plot of Natural3D result

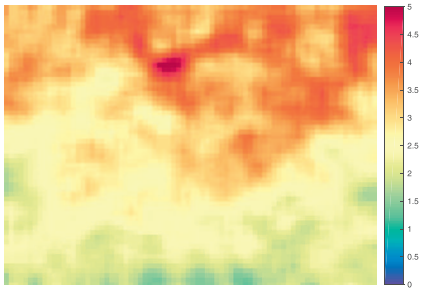
Figure 20. Example estimated depth maps along with the ground-truth depth map on the NYU Depth Dataset V2.

error slightly drops as the number of canonical depth patterns increases, the RMS value increases adversely. This result suggests that while it may be helpful to estimate relative distances between objects using more canonical depth patterns, the increased number of depth priors may result in inferior regression performance when estimating absolute distances. This result also agrees with our observation during the prior model development that five most common canonical depth patterns exist in natural environments. Therefore, using more than five clusters in the *k*-means algorithm may result in some redundant depth patterns, so the regression model of those redundant depth patterns will be trained with incomplete image data when estimating absolute distances, because the extracted image features belonging to similar depth patterns may be inaccurately

(a) Ground-truth depth map

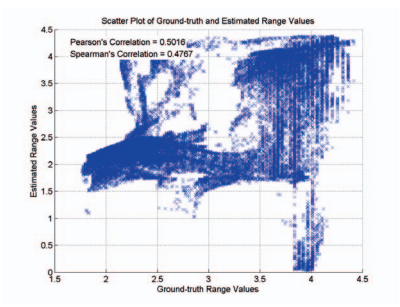


(b) Estimated depth map by Depth Transfer

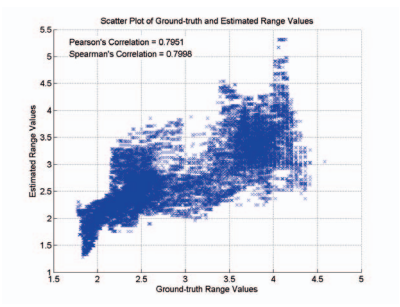


(c) Estimated depth map by Natural3D

(d) Natural image



(e) Scatter plot of Depth Transfer result



(f) Scatter plot of Natural3D result

Figure 21. Example estimated depth maps along with the ground-truth depth map on the NYU Depth Dataset V2.

Algorithm	Runtime per estimated depth map (s)
Depth Transfer	1490.53
Natural3D	161.05

Table 1. Computational complexity of monocular depth estimation algorithms.

classified into different clusters to train different regression models. As a result, to achieve the best depth estimation performance, we chose to use five canonical depth patterns: five clusters in the k-means algorithm, in our Natural3D implementation.

Regression model

A key component of Natural3D is the regression model used to estimate the mean range of each corresponding depth patch using the extracted depth-aware image feature vector $\mathbf{f}_I$. In our Natural3D implementation, we utilize a standard SVR model; in fact, the proposed Bayesian framework is generic enough to incorporate different types of regression models. To demonstrate this, we trained and tested two variants of Natural3D: the first using the general linear model regression, and the second with the generalized linear model (GLM) regression with the log link function and the assumption of a Poisson distribution, respectively. Note that the general linear model may be regarded as a special case of the GLM with the identity link function and the assumption of a normal (Gaussian) distribution. The parameter (mean) of both the Poisson and normal (Gaussian) distributions can be

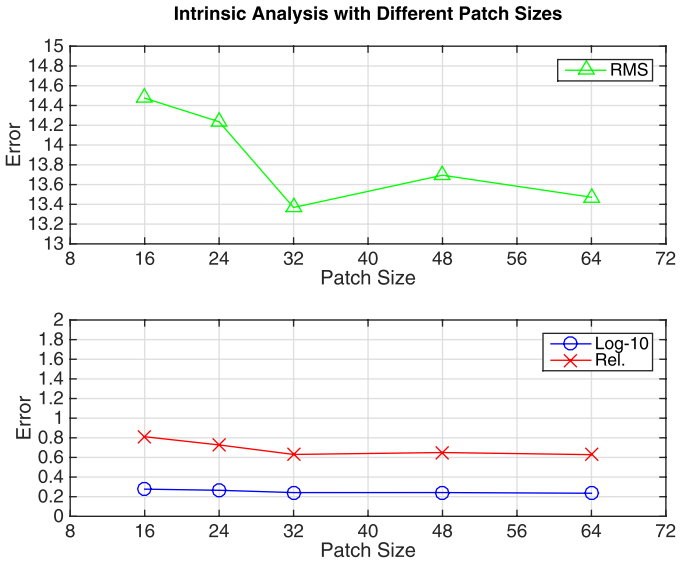


Figure 22. Plot of three error metrics as a function of patch size for Natural3D.

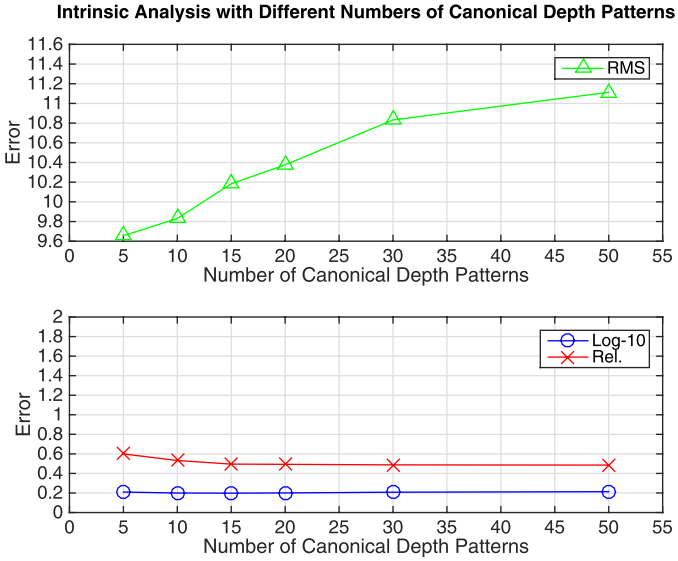


Figure 23. Plot of three error metrics as a function of the number of canonical depth patterns for Natural3D.

estimated using the NSS features extracted from the ground-truth depth patches and the associated image patches of the training data. We list the performance of these two variants in Table 2, which includes a baseline implementation of Natural3D as described herein. It can be seen that the two variants of Natural3D using linear regression models deliver comparable absolute depth estimation performance against the baseline in terms of the RMS metric. To attain the best performance when estimating both relative and absolute depths, we advocate a Natural3D implementation based on the standard SVR model, which can handle high-dimensional, depth-aware NSS features effectively.

Depth-aware NSS features

One of the major contributions of our work is the depth-aware features extracted using both established

	Metric		
Variant	Log10	Rel.	RMS
Baseline	0.2280	0.5964	12.9623
General linear model regression	0.2515	0.8069	13.6248
GLM regression with Poisson distribution	0.2505	0.8087	13.7498
HOG feature	0.2871	0.8653	16.9778
No Bayesian	0.2433	0.7287	15.5290

Table 2. Intrinsic analysis of Natural3D. *Notes.* Rel. = relative error, RMS = root mean square error, GLM = generalized linear model, HOG = histogram of oriented gradients.

and new bivariate and correlation NSS models. In order to demonstrate efficacy of these perceptually consistent NSS features, we trained and tested a variant of Natural3D using instead another type of highly popular image feature, the classic histogram of oriented gradients (HOG; Dalal & Triggs, 2005). This image feature has been widely used in computer vision to create low-level features for a plethora of different visual analysis tasks (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Felzenszwalb, Girshick, McAllester, & Ramanan, 2010), including object detection, classification, and recognition. We compared the depth estimation performance of the HOG variant with the Natural3D baseline implementation, as shown in Table 2. It can be seen that the HOG variant is not able to deliver comparable performance with the baseline in terms of all three error metrics, clearly demonstrating the effectiveness of the simple but relevant bivariate and correlation NSS features for depth estimation.

Bayesian inference

An essential ingredient of Natural3D is the Bayesian inference engine. To demonstrate the effectiveness of the priors and likelihoods Natural3D learns from natural images and registered depth maps, we implemented and compared with the baseline, a variant of Natural3D that simply trained a regression model to learn absolute depths directly from the extracted NSS features. As listed in the bottom row of Table 2, while the no-Bayesian implementation is able to deliver fair depth estimation performance, there does exist a noticeable drop in all three error metrics.

Conclusions

By exploiting reliable and robust statistical models describing the relationships between luminances and depths in natural scenes, we have created a perceptually relevant Bayesian model, called Natural3D, for recovering depth information from monocular (photographic) natural images. Two component models are learned from ground-truth depth maps: a prior model, including a dictionary of canonical depth patterns, and a likelihood model, which embeds co-occurrences of image and depth characteristics in natural scenes. As compared to several top-performing state-of-the-art computer vision methods, it delivers highly competitive performance in regards to estimating both absolute and relative depths from natural images. We also performed a thorough set of intrinsic analyses to acquire a better understanding of the contributions of the individual components of Natural3D to its performance, as well as the effectiveness of the extracted depth-aware NSS features.

The excellent performance attained by Natural3D implies that a biological visual system might be able to capture coarse depth estimates of the environment using the statistical information computed from retinal images at hand and the associations between image textures and true 3D geometric structures. We believe that the prior and likelihood models developed in Natural3D not only yield insights into how 3D structures in the environment might be recovered from image data, but could also be used to benefit a variety of 3D image/video and vision algorithms, such as creating fast estimates of scene or object depth in images from mobile camera devices. We envision that our future work will involve introducing deeper statistical models relating image and range data to recover more accurate and detailed depth information.

Keywords: depth estimation, Bayesian, human vision systems (HVS), natural scene statistics (NSS)

Acknowledgments

This research was supported by the National Science Foundation under Grants IIS-0917175 awarded to Lawrence K. Cormack, and IIS-1116656 awarded to Alan C. Bovik.

Commercial relationships: none.
Corresponding author: Che-Chun Su.
Email: ccsu@utexas.edu.
Address: Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA.

Footnotes

[1] By "natural scenes" we mean pictures of the real world, arising in both natural as well as man-made settings, obtained by a good quality camera under good (photopic) conditions without distortion.

[2] In the following sections, we will provide more details about the patch size $P \times P$ used in our implementation, and how performance is affected by different choices of P.

[3] There are no published results for Im2Depth and Eigen et al. (2014).

References

Baig, M. H., Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., & Sundaresan, N. (2014). Im2Depth: Scalable exemplar based depth transfer. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 145–152.

Bovik, A. (2013). Automatic prediction of perceptual image and video quality. *Proceedings of the IEEE*, *101*(9), 2008–2024.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3): 27, 1–27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Clark, M., & Bovik, A. C. (1989). Experiments in segmenting texton patterns using localized spatial filters. *Pattern Recognition*, *22*(6), 707–717.

Cobo-Lewis, A. B., & Yeh, Y.-Y. (1994). Selectivity of cyclopean masking for the spatial frequency of binocular disparity modulation. *Vision Research*, *34*(5), 607–620.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *1*, 886–893.

Delage, E., Lee, H., & Ng, A. (2006). A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *2*, 2418–2428.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Dosher, B., Sperling, G., & Wurst, S. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. *Vision Research*, *26*(6), 973–990.

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, *26*, 2366–2374.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Field, D. J. (1999). Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *357*(1760), 2527–2542.

Fouhey, D. F., Gupta, A., & Hebert, M. (2013). Data-driven 3D primitives for single image understanding. *Proceedings of the IEEE International Conference on Computer Vision*, 3392–3399.

Hassner, T., & Basri, R. (2006). Example based 3D reconstruction from single 2D images. *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, 15–22.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Hoiem, D., Efros, A. A., & Hebert, M. (2005). Automatic photo pop-up. *ACM Transactions on Graphics*, *24*(3), 577–584.

Intel Corporation. (2000). OpenCV: Camera calibration and 3D reconstruction. Available at http://docs.opencv.org/2.4.11/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html

Jordan, J. R., Geisler, W. S., & Bovik, A. C. (1990). Color as a source of information in the stereo correspondence process. *Vision Research*, *30*(12), 1955–1970.

Jou, J.-Y., & Bovik, A. C. (1989). Improved initial approximation and intensity-guided discontinuity detection in visible-surface reconstruction. *Computer Vision, Graphics, and Image Processing*, *47*(3), 292–326.

Karsch, K., Liu, C., & Kang, S. (2012). Depth extraction from video using non-parametric sampling. *Proceedings of the European Conference on Computer Vision*, *7576*, 775–788.

Ladický, L., Shi, J., & Pollefeys, M. (2014). Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 89–96.

Li, Q., & Wang, Z. (2009). Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal of Selected Topics in Signal Processing*, *3*(2), 202–211.

Lindeberg, T., & Garding, J. (1993). Shape from texture from a multi-scale perspective. In *Proceedings of the IEEE International Conference on Computer Vision*, 683–691.

Liu, B., Gould, S., & Koller, D. (2010). Single image depth estimation from predicted semantic labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1253–1260.

Liu, Y., Cormack, L. K., & Bovik, A. C. (2011). Statistical modeling of 3-D natural scenes with application to Bayesian stereopsis. *IEEE Transactions on Image Processing*, *20*(9), 2515–2530.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Lowe, D. (1999). Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, *2*, 1150–1157.

Lyu, S. (2011). Dependency reduction with divisive normalization: Justification and effectiveness. *Neural Computation*, *23*, 2942–2973.

Maki, A., Watanabe, M., & Wiles, C. (2002). Geotensity: Combining motion and lighting for 3D surface reconstruction. *International Journal of Computer Vision*, *48*(2), 75–90.

Malik, J., & Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *International Journal of Computer Vision*, *23*(2), 149–168.

Mallat, S. G. (1989a). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693.

Mallat, S. G. (1989b). Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(12), 2091–2110.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, *11*(2), 431–441.

Microsoft. (2010). Microsoft Kinect for Windows [Computer software]. Redmond, WA: Microsoft. Available at http://www.microsoft.com/en-us/kinectforwindows/

Moorthy, A. K., & Bovik, A. C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, *20*(12), 3350–3364.

Nagai, T., Naruse, T., Ikehara, M., & Kurematsu, A. (2002). HMM-based surface reconstruction from single images. *Proceedings of the IEEE International Conference on Image Processing*, *2*, 561–564.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Nerual Systems*, *7*(2), 333–339.

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, *17*(8), 1665–1699.

Owens, A., Xiao, J., Torralba, A., & Freeman, W. (2013). Shape anchors for data-driven multi-view reconstruction. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 33–40). Piscataway, NJ: IEEE Publishing.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Portilla, J., Strela, V., Wainwright, M. J., & Simoncelli, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, *12*(11), 1338–1351.

Potetz, B., & Lee, T. S. (2003). Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America A*, *20*(7), 1292–1303.

Potetz, B., & Lee, T. S. (2006). Scaling laws in natural scenes and the inference of 3D shape. *Advances in Neural Information Processing Systems*, *18*, 1089–1096.

Rajashekar, U., Wang, Z., & Simoncelli, E. P. (2010). Perceptual quality assessment of color images using adaptive signal representation. In B. E. Rogowitz and T. N. Pappas (Eds.), *SPIE International Conference on Human Vision and Electronic Imaging XV*, Vol. 7527.

RIEGL Laser Measurement Systems. (2009). RIEGL VZ-400 3D Terrestrial Laser Scanner. Available at http://products.rieglusa.com/product/terrestrial-scanners/vz-400-3d-laser-scanners.

Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*(4), 517–548.

Saxena, A., Chung, S. H., & Ng, A. Y. (2005). Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, *17*, 1161–1168.

Saxena, A., Sun, M., & Ng, A. Y. (2005). Make3D Laser+Image Dataset-1. Available at http://make3d.cs.cornell.edu/data.html

Saxena, A., Sun, M., & Ng, A. (2009). Make3D:

Learning 3D scene structure from a single still images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 824–840.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207–1245.

Schumer, R. A., & Ganz, L. (1979). Independent stereoscopic channels for different extents of spatial pooling. *Vision Research*, *19*, 1303–1314.

Schwartz, B. J., & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society*, *21*(6), 456–458.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*, 819–825.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Sharifi, K., & Leon-Garcia, A. (1995). Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, *5*(1), 52–56.

Sheikh, H., & Bovik, A. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, *15*(2), 430–444.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Proceedings of the European conference on computer vision* (Vol. 5, pp. 746–760). Berlin, Heidelberg: Springer-Verlag.

Silberman, N., Kohli, P., Hoiem, D., & Fergus, R. (2012). NYU Depth Dataset V2 [Data file]. Available at http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

Simoncelli, E. P. (1999). Modeling the joint statistics of images in the wavelet domain. *Proceedings of SPIE, Wavelet Applications in Signal and Image Processing VII*, *3813*, 188–195.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *IEEE International Conference on Image Processing*, *3*, 444–447.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Sinno, Z., & Bovik, A. C. (2015). Generalizing a closed-form correlation model of oriented bandpass natural images. *IEEE Global Conference on Signal and Information Processing*, 373–377.

Su, C.-C. (2016). Software release: Natural3D [Computer software]. Available at http://live.ece.utexas.edu/research/3dnss/index.html

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2013). Color and depth priors in natural images. *IEEE Transactions on Image Processing*, *22*(6), 2259–2274.

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2014a). Bivariate statistical modeling of color and range in natural scenes. *Proceedings of SPIE, Human Vision and Electronic Imaging XIX*, 9014.

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2014b). New bivariate statistical model of natural image correlations. *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 5362–5366.

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2015a). Closed-form correlation model of oriented bandpass natural images. *IEEE Signal Processing Letters*, *22*(1), 21–25.

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2015b). Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation. *IEEE Transactions on Image Processing*, *24*(5), 1685–1699.

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2016a). Experimental results of monocular depth estimation algorithms. Available at http://live.ece.utexas.edu/research/3dnss/index.html

Su, C.-C., Cormack, L. K., & Bovik, A. C. (2016b). LIVE Color+3D Database Release-2 [Database]. Available at http://live.ece.utexas.edu/research/3dnss/live_color_plus_3d.html

Tang, H., Joshi, N., & Kapoor, A. (2011). Learning a blind measure of perceptual image quality. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 305–312.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(24), 1226–1238.

Tyler, C. W. (1974). Depth perception in disparity gratings. *Nature*, *251*, 140–142.

Tyler, C. W. (1975). Spatial organization of binocular disparity sensitivity. *Vision Research*, *15*(5), 583–590.

Tyler, C. W. (1983). Sensory processing of binocular disparity. *Vergence eye movements: Basic and*

clinical aspects (pp. 199–295). Oxford, UK: Butterworth-Heinemann.

Wainwright, M. J., Schwartz, O., & Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In R. Rao, B. Olshausen, & M. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (p. 203–222). Cambridge, MA: MIT Press.

Wang, Z., & Bovik, A. C. (2011). Reduced- and no-reference image quality assessment: The natural scene statistic model approach. *IEEE Signal Processing Magazine*, *28*(6), 29–40.

Zhang, R., Tsai, P.-S., Cryer, J. E., & Shah, M. (1999). Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(8), 690–706.

Zhang, X., & Wandell, B. A. (1997). A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display*, *5*(1), 61–63.