Bayesian Depth Estimation from Monocular Natural Images

Che-Chun Su, *Student Member, IEEE,* Lawrence K. Cormack, and Alan C. Bovik, *Fellow, IEEE*

Abstract

We consider the problem of estimating a dense depth map from a single monocular natural image. Inspired by psychophysical evidence of visual signal processing in human vision systems (HVS), we propose a Bayesian framework to recover detailed 3D scene structures by exploiting reliable and robust natural scene statistics (NSS) models of natural images and depth maps. Specifically, we utilize the statistical relationships between local image features and depth variations inherent in natural images. By observing that similar depth structures may exist in different types of luminance textured regions in natural scenes, we build a dictionary of canonical depth patterns as the prior, and fit a multivariate Gaussian mixture (MGM) model to associate local image features to different depth patterns as the likelihood. Compared with the state-of-the-art depth estimation method, we achieve superior performance in terms of pixel-wise estimated depth error, but better capability of recovering relative distant relationships between different objects and regions in natural images.

Index Terms

Depth estimation, Bayesian, human vision systems (HVS), natural scene statistics (NSS)

- *Che-Chun Su and Alan C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA.*
E-mail: ccsu@utexas.edu; bovik@ece.utexas.edu

1 Introduction

With the rapid growth and wide spread popularity of 3D films and entertainment devices, understanding how depth information and 2D image data interact in natural images and videos has become of more importance in the past several years. In particular, recovering the three-dimensional structure of a scene from a single monocular image has been regarded as a fundamental problem in image/video processing and computer vision disciplines. More accurate depth estimation results in better understanding of the geometric relationships between objects in natural images, which would be beneficial to various image/video and vision applications, e.g., robotics, surveillance, scene understanding, perception quality, etc.

By seamlessly combining both binocular and monocular cues, humans are able to perceive depth and reconstruct the geometry of the 3D visual space so quickly and effortlessly that an individual rarely feels how difficult and ill-posed this problem can be. Even either given a single color image or with one eye closed, the human vision system can still acquire accurate depth structures of natural environments and relative distant relationships between different objects. However, for computer programs and robotics, estimating range (egocentric distance) from a single monocular image has been known as a very difficult problem generally approached by using a combination of well-known depth cues, e.g., color, texture, perspective, etc.

Much work on 3D scene reconstruction has focused on binocular vision, i.e., stereopsis. In [1], Scharstein and Szeliski provided a thorough review and summary of dense two-frame stereo algorithms. Many other depth recovering algorithms require multiple images, including structure from motion [2] and depth from defocus [3]. These algorithms consider only the geometric/triangulation

- *Lawrence K. Cormack is with the Department of Psychology, The University of Texas at Austin, Austin, TX 78712, USA.*
  E-mail: cormack@utexas.edu

differences, while there is also a variety of monocular cues that contain useful and depth information.

Recently, there have been many different methods and algorithms developed to tackle the problem of depth estimation from a single monocular image. Examples include shape from shading [4], [5] and shape from texture [6], [7]; however, it is difficult to apply these algorithms to surfaces without fairly uniform texture and luminance variations. Nagai *et al.* [8] used Hidden Markov Models (HMM) to reconstruct surfaces of known, fixed objects such as hands and faces from single images. In [9], an example-based approach was proposed by Hassner *et al.* to estimate the depth of an object given some known classified category.

One of the first methods that utilizes monocular image features, proposed by Hoiem *et al.* [10], reconstructs a simple 3D model of outdoor scenes by making the assumption that an image could be divided into a few planar surfaces, and pixels could be classified into limited labels, e.g., ground, sky, and vertical walls. Delage *et al.* [11] 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. In [12], [13], a supervised learning strategy was devised by Saxena *et al.* to infer absolute depth of each pixel in the monocular image. They assumed that most 3D scenes are made up of many small, approximately planar surfaces, and use a Markov Random Field (MRF) to model both the monocular depth cues, e.g., texture variations and gradients, as well as the relationships between different parts of the image. In [14], Torralba and Oliva studied the relationship between the Fourier spectrum of an image and its mean depth. Specifically, they proposed a probability model to estimate absolute mean depth of a scene using both the global and local spectral signatures of an natural image. In [15], Liu *et al.* incorporated semantic labels to guide the 3D reconstruction process, and achieve better depth estimates of each pixel in a scene. By conditioning on different semantic labels, they were able to better model the absolute depth as a function of local pixel appearance.

Recently, Karsch *et al.* [16] presented an optimization framework to generate the most likely depth map by first matching high-level image features to find candidates from the database, and then warping those candidate depth maps with spatial regularization constraints.

Natural scene statistics (NSS) have proven to provide abundant and useful resources towards both understanding the evolution of human vision systems (HVS) [17], [18] and solving diverse image/video and vision problems [19]–[22]. There has also been work conducted on exploring 3D NSS and their applications. For example, Potetz *et al.* [23] examined the relationships between luminance and range over multiple scales and applied their results to shape-from-shading problems. Yang *et al.* [24] explored the statistical relationships between luminance and disparity in the wavelet domain, and applied the derived models to a Bayesian stereo algorithm. In [25], Su *et al.* proposed reliable statistical models for both marginal and conditional distributions of luminance/chrominance and disparity in natural images, and used these models to significantly improve a chromatic Bayesian stereo algorithm. Recently, Su *et al.* developed new bivariate and correlation NSS models that well capture the higher-order dependencies between spatially adjacent bandpass responses in both natural images and depth maps [26], [27]. In [28], the authors further utilized these robust models to propose a generic quality evaluation framework on stereoscopic image pairs with superior performance to state-of-the-art algorithms.

In this work, inspired by psychophysical evidence of visual signal processing in HVS, we propose a Bayesian framework for estimating depth from single monocular images by exploiting reliable and robust NSS models of natural images and depth maps [26], [27]. The proposed Bayesian framework is trained and tested on an accurately co-registered database of natural image and range data, the LIVE 3D+Color Database - Release 2 [29], which consists of 99 pairs of natural images and ground-truth depth maps in high-definition resolution of 1920×1080.

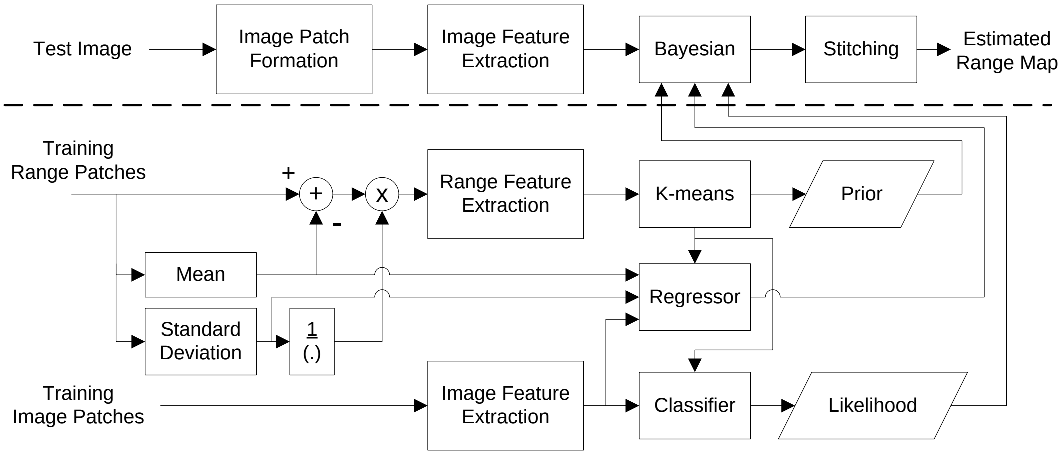


Fig. 1. Block diagram of the proposed Bayesian framework for depth estimation.

The rest of this paper is organized as follows. Section 2 describes details of the proposed Bayesian framework, including the prior and likelihood models. The experimental results are presented in Section 3, followed by the conclusions in Section 4.

2 Proposed Bayesian Framework

Figure 1 shows a block diagram of the proposed Bayesian framework for depth estimation from monocular natural images. The framework is divided into two parts, training and testing. For the training part, we first collect patches of size 32x32 from both natural images and corresponding ground-truth depth maps, and then extract natural scene statistical features for each patch pair. Next, to explore the embedded depth information in natural images, we learn the priors and likelihoods from these perceptual image and depth features. For the testing part, an input image is divided into overlapping 32x32 patches, and the same set of features is extracted from each image patch. The corresponding depth patch is estimated for each image patch using a Bayesian model with the learned priors and likelihoods. Finally, all the depth patches are stitched together to form the estimated depth map. The details of each component of the proposed Bayesian framework are explained in the following subsections.

2.1 Perceptual Decomposition

Human vision systems (HVS) extract abundant information from natural environments by processing visual stimuli through different levels of decomposition and interpretation. By emulating how HVS process natural image and depth information, a variety of statistical models have been proposed to fit the bandpass responses of luminance/chrominance and depth/disparity in natural scenes [24], [25], [30]. In this work, since we want to learn and exploit the statistics relating depth perception to natural images, we apply certain perceptually relevant pre-processing steps on natural image luminance, and extract depth-aware features from both univariate and bivariate empirical response distributions.

We acquire luminance by transforming pristine color images into the perceptually relevant CIELAB color space, which is optimized to quantify perceptual color differences and better corresponds to human color perception than does the perceptually nonuniform RGB space [31]. Each luminance image (L*) is then transformed by the steerable pyramid decomposition, which is an over-complete wavelet transform that allows for increased orientation selectivity [32]. 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 [33], [34]. Specifically, in the implementation of the proposed Bayesian framework, we utilize the steerable pyramid decomposition with five scales, indexed from 1 (finest) to 5 (coarsest), and twelve frequency-tuning orientations: $0, \frac{1}{12}\pi, \ldots, \frac{11}{12}\pi$.

After applying the multi-scale, multi-orientation decomposition, we perform the perceptually significant process of divisive normalization on the luminance wavelet coefficients of all of the sub-bands [35]. Divisive normalization, i.e., sensory gain control, was proposed in the psychophysical literature to account for the nonlinear behavior of human perceptual neurons [36]. The divisive normalization transform (DNT) used in our work is implemented as follows [37]:

$$\begin{aligned} u(x_i, y_i) &= \frac{w(x_i, y_i)}{\sqrt{s + \mathbf{w}_\mathbf{g}^\top \mathbf{w}_\mathbf{g}}} \\ &= \frac{w(x_i, y_i)}{\sqrt{s + \sum_j g(x_j, y_j) w(x_j, y_j)^2}} \end{aligned} \tag{1}$$

where (x_i, y_i) are spatial coordinates, w are the wavelet coefficients, u are the coefficients after DNT, s is a semi-saturation constant, the weighted sum occurs over neighborhood pixels indexed by j, and $\{g(x_j, y_j)\}$ is a finite-extent Gaussian weighting function.

In the following subsections, we explain the details of extracting both image and depth features from these divisively normalized sub-band responses to learn the prior and likelihood for depth estimation.

2.2 Image Feature Extraction

It has been known that there exist statistical relationships between luminance intensity and depth information in natural scenes [30], and a variety of univariate statistical models have been proposed to fit the bandpass responses of luminance/chrominance and disparity [24], [25]. Recently, new bivariate and correlation statistical models have been developed to capture spatial dependencies between neighboring sub-band responses in natural images [27]. In the proposed Bayesian framework, we exploit these natural scene statistical features to learn the relationships between the projected image luminance and the embedded depth information in natural environments.

2.2.1 Univariate NSS Feature

Considerable work has been conducted on modeling the statistics of natural images using multi-scale, multi-orientation transforms, e.g., Gabor filters, wavelets, etc [18], [38]. Here we use the univariate generalized Gaussian distribution (GGD) to fit the empirical histograms of luminance sub-band responses, i.e., u in Eq. (1), of each image patch. The probability density function of a univariate GGD with zero mean is:

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

where $\Gamma(\cdot)$ is the ordinary gamma function and α_u and β_u are scale and shape parameters, respectively. The two resulting GGD parameters from each sub-band, scale and shape, are included in the feature set of each image patch.

2.2.2 Bivariate NSS Feature

In addition to univariate statistics in natural luminance, we further exploit higher-order dependencies that exist between spatially neighboring bandpass image responses. Specifically, we examine the bivariate distributions of horizontally adjacent sub-band responses, which are sampled from locations (x, y) and $(x+1, y)$ in an image patch. To model these empirical joint histograms, we utilize a multivariate generalized Gaussian distribution (MGGD), which include both the multivariate Gaussian and Laplace distributions as special cases. The probability density function of a multivariate generalized Gaussian distribution is defined as:

$$p(\mathbf{x}; \mathbf{M}, \alpha_b, \beta_b) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\alpha_b, \beta_b}(\mathbf{x}^\top \mathbf{M}^{-1} \mathbf{x}) \tag{3}$$

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

$$g_{\alpha_b,\beta_b}(y) = \frac{\beta_b \Gamma(\frac{N}{2})}{(2^{\frac{1}{\beta_b}} \pi \alpha_b)^{\frac{N}{2}} \Gamma(\frac{N}{2\beta_b})} e^{-\frac{1}{2}(\frac{y}{\alpha_b})^{\beta_b}} \tag{4}$$

where $y \in \mathbb{R}^+$. Note that when $\beta_b = 0.5$, Eq. (3) becomes the multivariate Laplacian distribution, and when $\beta_b = 1$, Eq. (3) corresponds to the multivariate Gaussian distribution. Moreover, when $\beta_b \to \infty$, the MGGD converges to a multivariate uniform distribution. In our implementation, we model the bivariate empirical histograms of horizontally adjacent sub-band coefficients of each image patch using a bivariate generalized Gaussian distribution (BGGD) with $N = 2$ in Eq. (3). The parameters of the BGGD can be obtained on the bandpass coefficients of image patches using the maximum likelihood estimator (MLE) algorithm described in [26]. We include both the scale and shape parameters, α_b and β_b, in the image patch feature set.

2.2.3 Correlation NSS Feature

In addition to the univariate and bivariate GGD models fitting empirical distributions of sub-band coefficients in natural images, there exist higher-order dependencies that have not been utilized between spatially neighboring bandpass luminance responses. In particular, we have found that the correlation coefficients between spatially adjacent bandpass responses posses strong orientation dependencies [27]. For example, the horizontally adjacent bandpass responses are most correlated when the sub-band tuning orientation aligns at $\frac{1}{2}\pi$, and become nearly uncorrelated at orientation 0 (rad) and π, indicating its periodicity with relative orientation of spatial and sub-band tuning orientation. This relative orientation regularity on correlation coefficients between spatial neighboring sub-band responses provides useful cues on distinguishing areas in the image with different geometric structure, e.g., depth discontinuities, smooth surfaces, etc.

We have found that the periodic relative orientation dependency of the correlation coefficients between spatially adjacent sub-band responses can be well modeled as an exponentiated sine function given by:

$$y = f(x_1, x_2) = A \left[\frac{1 + \sin\left(\frac{2\pi(x_2 - x_1)}{T} + \theta\right)}{2} \right]^{\gamma} + c \tag{5}$$

where y is the correlation coefficients between spatially adjacent bandpass responses, x_1 and x_2 represent spatial and sub-band tuning orientations, respectively, A is the amplitude, T is the period, θ is the phase, γ is the exponent, and c

is the offset. The correlation coefficient is periodic with π of relative orientation and reaches maximum when $|x_2 - x_1| = k \cdot \frac{\pi}{2}, k \in \mathbb{N}$, yielding a three-parameter exponentiated sine model with amplitude A, exponent γ, and offset c, by fixing $T = \pi$ and $\theta = \frac{\pi}{2}$:

$$\begin{aligned} y = f(x_1, x_2) &= A \left[\frac{1 + \cos\left(2(x_2 - x_1)\right)}{2} \right]^{\gamma} + c \\ &= A \left[\cos\left(x_2 - x_1\right) \right]^{2\gamma} + c \end{aligned} \tag{6}$$

In our implementation, we compute the correlation coefficients between horizontally adjacent sub-band responses on each image patch at different scales, fit the exponentiated sine model, and include all three parameters, A, γ, and c, into the feature set.

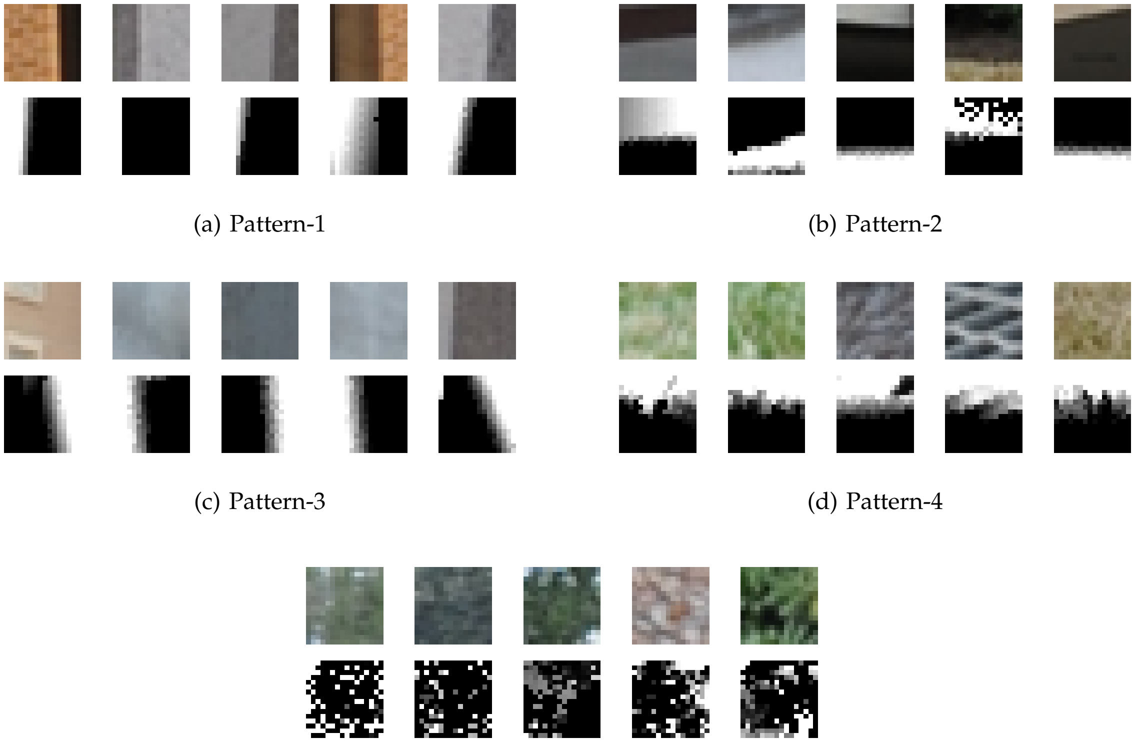
As a result, the depth-aware feature vector $\mathbf{f}$ that we use to characterize each image patch is formed as:

$$\mathbf{f} = \left[\{\alpha_{u,k}\}, \{\beta_{u,k}\}, \{\alpha_{b,k}\}, \{\beta_{b,k}\}, \{A_s\}, \{\gamma_s\}, \{c_s\} \right]^{\top} \tag{7}$$

where $k \in \{1, \ldots, K\}$, K is the number of sub-bands, and $s \in \{1, \ldots, S\}$, S is the number of scales.

2.3 Depth Feature Extraction

Unlike luminance intensity in natural images, range/depth maps captured from natural environments tend to possess smooth surfaces with relatively few textures. Based on this observation, we use the histogram of gradient magnitudes as depth features to characterize different types of depth patches extracted from ground-truth depth maps [39]. As shown in Fig. 1, we first subtract the mean from each depth patch, and divide the difference by its standard deviation to obtain a normalized depth patch. Then, we compute the depth gradient magnitudes at each pixel location, and obtain the corresponding histogram on the entire patch. Note that the histogram of depth gradient magnitudes is computed along eight canonical orientations, i.e., 0 (rad), $\frac{1}{8}\pi$, … , $\frac{7}{8}\pi$, resulting in an eight-bin histogram for each depth patch.



(a) Pattern-1

(b) Pattern-2

(c) Pattern-3

(d) Pattern-4

(e) Pattern-5

Fig. 2. Examples of different canonical depth patterns.

In addition to the gradient histogram, we also compute the histogram of bandpass response magnitudes on the perceptually decomposed depth patches. Specifically, we compute the divisively normalized wavelet responses u using Eq. (1) at each pixel location, and obtain a histogram by binning the responses along different sub-band tuning orientations. The same eight canonical orientations as in the gradient magnitude histogram are used for the bandpass response histogram. Therefore, we end up with a 16-dimension feature vector to characterize each depth patch for learning the prior and likelihood in the proposed Bayesian framework.

2.4 Prior

It may be observed that in natural images, discontinuities in depth maps usually implies that there are luminance edges at the same location in the corresponding color images. In other words, patches of different luminance appearance and texture may have similar depth patterns, due to the common geometric structures [24]. Moreover, without effects of ambient light and textured surfaces, depth maps tend to posses simpler, more regular patterns than natural luminance images. Therefore, inspired by these observations, we build a dictionary of canonical depth patterns by clustering the features extracted from depth patches with the k-means algorithm.

Figure 2 shows some examples of different canonical depth patterns extracted by the k-means algorithm with five clusters. For each canonical depth pattern, the top row shows the clustered depth patches using the extracted features, and the bottom row the corresponding image patches. We can see that these canonical depth patterns contain different types of geometric structures, including depth discontinuity along the horizontal direction (pattern-1), depth discontinuity along the vertical direction (pattern-2), smooth variation of depth along the horizontal direction (pattern-3), smooth variation of depth along the vertical direction (pattern-4), and a busy, complicated pattern of depth changes (pattern-5). In fact, we can easily find these canonical range patterns in natural scenes, e.g., most of the busy, complicated range patterns appear in areas filled with tree leaves and grass. In addition, as the number of clusters used in the k-means algorithm increases, these five canonical depth patterns still exist, while different clusters of range patches may share similar structures. As a result, the depth prior of the proposed Bayesian framework consists of the normalized residual depth patches of each canonical depth pattern, as well as the portion of depth patches belonging to each pattern among all depth patches, i.e., $p(n)$, where $n \in \{1, \ldots, N\}$ and N is the number of canonical depth patterns, i.e., the number of clusters used in the k-means algorithm.

2.5 Likelihood

As we can observe from the canonical depth patterns shown in Fig. 2, the depth discontinuities in range maps consistently match the luminance edges in natural images, while some textured areas with variations in luminance/chrominance may not necessarily correspond to depth changes, resulting in smooth surfaces with low gradients in range maps. In other words, there exist high correlations between image edges and depth discontinuities. For example, if there are strong variations in a natural image, i.e., large bandpass responses, there is a high likelihood of co-located variations, i.e. large depth gradients, in the corresponding range map. To better utilize these relationships between image and depth variations in natural environments, we derive a likelihood model which aims to associate image patches to different canonical depth patterns.

First, assume we obtain N canonical depth patterns from the prior using the k-means clustering algorithm. Then, we assign each image patch a label indicating the canonical depth pattern of its corresponding depth patch. Based on these labeling results, we use the depth-aware feature set, i.e., $\mathbf{f}$ in Eq. 7, extracted from each image patch to train a classifier using a multivariate Gaussian mixture (MGM) model. The reason that the MGM model fits well to this classification task is that, as observed in Fig. 2, image patches with different appearances and/or textured surfaces may possess the same canonical depth pattern due to the similar underlying geometric structures. Therefore, we can take advantage of the Gaussian mixture model trained for each canonical depth pattern to handle the heterogeneity of its image patches. The MGM model for the n-th canonical range pattern is given by

$$p(\mathbf{x};\theta_n) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{8}$$

where $n \in \{1, \ldots, N\}$, θ_n is the model parameter, $\mathbf{x}$ is a multi-dimensional 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 parametrized by $\theta_n = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, m \in \{1, \ldots, M\}$, 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

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_m|^{1/2}} e^{\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^\top \boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]} \quad (9)$$

where D is the dimensionality of $\mathbf{x}$. Here we use the depth-aware feature set $\mathbf{f}$ as $\mathbf{x} \in \mathbb{R}^D$. Therefore, the likelihood probability of seeing an image patch with the extracted feature $\mathbf{f}$ given a particular canonical depth pattern n can be expressed as $p(\mathbf{f};\theta_n)$.

2.6 Regression on Mean Depth

As discussed in Sec. 2.3 and 2.4, before we extract features from depth patches to learn the prior, we normalize each depth patch by removing the mean and standard deviation to better characterize the depth patterns. To add the mean back to each depth patch when estimating true range values of the test image patches, we need to learn a mapping from the image feature space using a regression model. In other words, given an input image patch, we can utilize the trained regressor to estimate the mean range of the corresponding depth patch using the extracted depth-aware image feature set $\mathbf{f}$. In addition to $\mathbf{f}$, we exploit another important monocular depth cues in natural environments to recover the true range. In [30], it has been demonstrated that there is general correlation between the brightness and the distance in natural scenes. Here we utilize this "the brighter the nearer" correlation to estimate the mean range value from the average luminance intensity of the image patch. Moreover, it can be observed that in natural images, the distance from the nodal point to any point in the scene generally increases 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. Therefore, we introduce one more feature into the regressor on mean depth values, the normalized y-coordinate of each patch in the image, given by

$$f_y = \frac{p_y}{I_h} \tag{10}$$

where p_y is the y-coordinate of the image patch, and I_h is the height of the image. As a result, the feature of the image patch used in the regression model to learn the mean depth value includes the depth-aware feature set $\mathbf{f}$, the average luminance intensity, as well as the normalized y-coordinate f_y. In the proposed Bayesian framework, we adopt the support vector regression (SVR) model to serve as the regressor. The proposed Bayesian framework is generically amenable to the application of any kind of regressor. Our implementation utilizes a support vector machine (SVM) regressor (SVR) [40] using multiple train-test sets as described in Sec. 3. SVR is generally noted for being able to handle high dimensional data [41]. We implement the SVR model with a radial basis function (RBF) kernel using the LIBSVM package [42].

2.7 Bayesian Model

The primary component of the proposed framework is the Bayesian model that incorporates the prior of canonical depth patterns, the likelihood associating image patches to different canonical depth patterns, and the regression model recovering the mean range values for each image patch. Given a test image, we first divide it into overlapped patches of size 32x32 as in the training phase, where a $\frac{1}{4}$ overlap is used, i.e., patches overlap each other by 8 pixels along both dimensions. Next, the depth-aware feature vector $\mathbf{f}$ is extracted from each image patch, as well as the average luminance intensity and the normalized y-coordinate for mean depth regression. Then, the extracted feature $\mathbf{f}$ is fed into the trained prior, likelihood, and regression models to form a Bayesian inference of the corresponding estimated depth patch. In particular, the estimated depth patch $\mathbf{D}$ of an image patch is formed as follows:

$$\mathbf{D} = \mathbf{D}_{r,n} + \mu_n \tag{11}$$

where $\mathbf{D}_{r,n}$ is the normalized residual depth patch of the estimated canonical depth pattern n, μ_n is the corresponding mean depth value obtained from the regression model, and n represents the estimated canonical depth pattern derived from the prior and likelihood, which is given by:

$$\begin{aligned} n &= \underset{n'}{\operatorname{argmax}}\,\{p(n'|\mathbf{f})\} = \underset{n'}{\operatorname{argmax}}\,\{p(\mathbf{f}|n')p(n')\} \\ &= \underset{n'}{\operatorname{argmax}}\,\{p(\mathbf{f};\theta_{n'})p(n')\} \end{aligned} \tag{12}$$

where $p(\mathbf{f}|n') = p(\mathbf{f};\theta_{n'})$ is the likelihood probability (Eq. (8)) of seeing an image patch with the extracted feature $\mathbf{f}$ given a canonical depth pattern n' as derived in Sec. 2.5, and $p(n')$ is the corresponding prior probability of n' as obtained in Sec. 2.4.

2.8 Stitching

Finally, the last stage of the proposed Bayesian framework is to stitch all depth patches together to form the final estimated depth map of the input test image. Note that the stitching operation used here is simply averaging the estimated range values of the overlapped pixels between depth patches.

3 Experimental Results

To evaluate the performance of the proposed depth estimation framework, we trained and tested the Bayesian model on the LIVE Color+3D Database - Release 2 [29], which consists of 99 pairs of color images and ground-truth range maps in high-definition resolution of 1920×1080. The dense, accurately co-registered depth maps in this database provide rich information regarding natural depth statistics, and is also an excellent resource for evaluating depth estimation algorithms. To avoid overlap between training and testing image/depth content, we split the whole database into 80% training and 20% testing subsets at each train-test iteration. This train-test procedure was repeated 50 times to ensure that there was no bias introduced due to the image/depth content used for

TABLE 1

Performance Comparison of Different Depth Estimation Algorithms (Median across Train-Test Splits)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.4196	0.5197	0.6399	13.0671
Proposed Bayesian	0.4404	0.5654	0.5969	12.8417

training. We compared the proposed Bayesian framework with a state-of-the-art depth estimation method, Depth Transfer [16].

Table 1 and 2 show the comparison results in terms of four different error metrics. First, we report the two common metrics, the relative error (Rel.):

$$\sum_{i=1}^{I} \frac{|\mathbf{D}(x_i, y_i) - \mathbf{D}^*(x_i, y_i)| / \mathbf{D}^*(x_i, y_i)}{I} \tag{13}$$

and the root mean squared error (RMS):

$$\sqrt{\sum_{i=1}^{I} \frac{[\mathbf{D}(x_i, y_i) - \mathbf{D}^*(x_i, y_i)]^2}{I}} \tag{14}$$

where $\mathbf{D}(x_i, y_i)$ and $\mathbf{D}^*(x_i, y_i)$ represent the estimated and ground-truth depth map at pixel location (x_i, y_i), respectively, and I is the number of pixels. In addition, to examine how well a depth estimation method is able to recover the relative distance in natural scenes, we report two different correlation coefficients between the estimated and ground-truth depth values, the Pearson's linear correlation coefficient ρ_p and the Spearman's rank order correlation coefficient ρ_s. Specifically, ρ_p and ρ_s measure the accuracy and monotonicity, respectively, of the estimated range values by a depth estimation algorithm against the ground-truth range values, where a value of 1 indicates perfect correlation.

As we can see from Table 1, which shows the median metric performance across train-test splits, the proposed Bayesian framework outperforms Depth

TABLE 2

Performance Comparison of Different Depth Estimation Algorithms (Standard Deviation across Train-Test Splits)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.2205	0.2461	0.3891	8.3052
Proposed Bayesian	0.1987	0.2253	0.3858	8.3921

TABLE 3

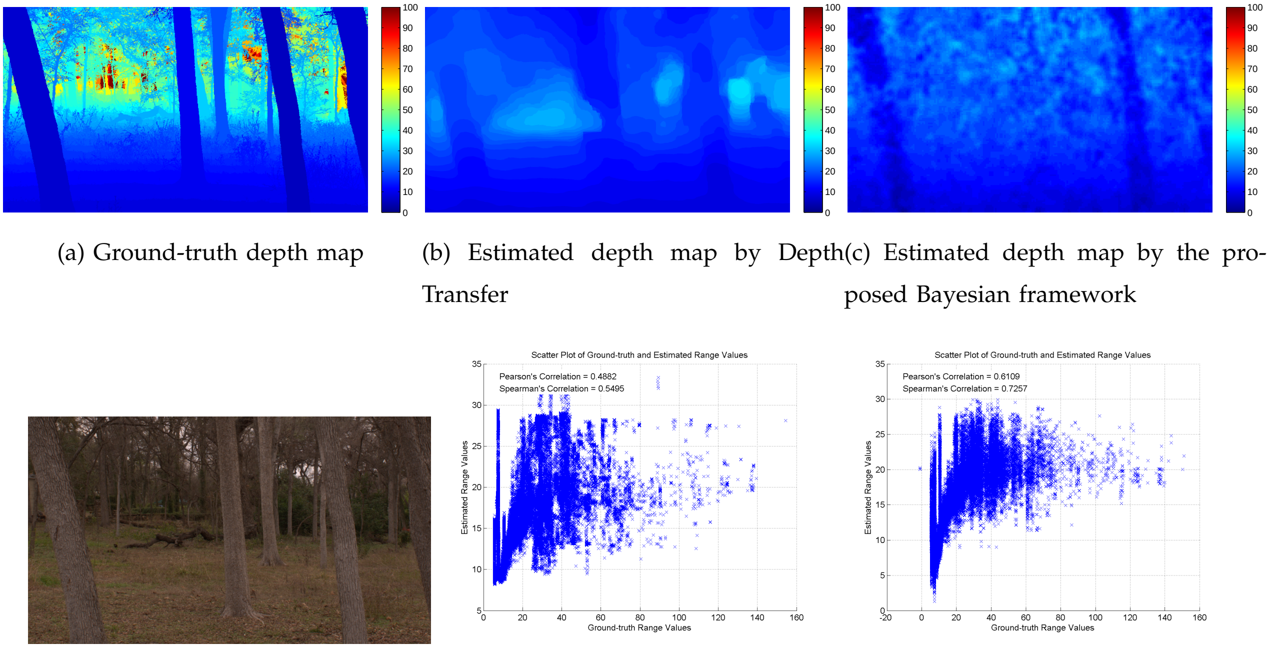
Computational Complexity of Different Depth Estimation Algorithms

Algorithm	Runtime per Estimated Depth Map (s)
Depth Transfer	1490.53
Proposed Bayesian	161.05

Transfer in terms of all four error metrics. The higher correlation performance of the proposed Bayesian framework indicates that it is capable of recovering more accurate relative distances between different objects and regions in natural scenes. In addition, the proposed Bayesian framework achieves both lower relative and RMS errors than Depth Transfer, meaning that the depth maps estimated by the proposed Bayesian framework are closer to the ground-truth values. Table 2 shows the standard deviation of different error metrics across train-test splits, which signifies the performance consistency of the examined depth estimation algorithms. It can be seen that the proposed Bayesian framework delivers more consistent performance in terms of both linear and rank order correlation coefficients, while providing similar performance consistency on estimating absolute depth as Depth Transfer.

In addition to the quantitative comparison, we also give a visual comparison by showing examples of the estimated depth maps from the two examined depth estimation algorithms along with the corresponding ground-truth range maps, as demonstrated in Figures 3 to 6. We also draw the scatter plots between the estimated and the ground-truth range values to gain a broader perspective of performance. In general, we can see that Depth Transfer tends to over-smooth the estimated range maps due to its smoothness constraint, while the proposed Bayesian framework is able to discover more detailed structures in the scene. For example, in Fig. 4, Depth Transfer is not able to capture the tree trunks in the foreground, and it incorrectly mix the tree trunks with the background. On the other hand, the proposed Bayesian framework better outlines the tree trunks, achieving both higher linear and rank order correlations against the ground-truth depth map. In Fig. 5, Scene-3 contains a mixture of human objects and a tree branch, posing more challenging content for depth estimation algorithms. The proposed Bayesian framework successfully captures the intersection of the human hand and the tree branch, while Depth Transfer fails to recover this complicated structure by smoothing out the tree branch into the background. Moreover, it can be clearly seen in Fig. 6 that the proposed Bayesian framework reconstructs the main tree structures, while Depth Transfer incorrectly combines two separate tree trunks into one. Finally, the correlation coefficients shown in all figures also match the numerical results, confirming that the proposed Bayesian framework achieves superior performance at recovering relative distances in natural scenes. Note that there is no smoothness constraint in the proposed Bayesian framework, where only simple averaging operation is performed on overlapped pixels between patches. This implies some improvement can be readily made for the proposed Bayesian framework.

Another advantage of the proposed Bayesian framework is no use of iteration, resulting in much less computational complexity. Table 3 shows the runtime per estimated depth map for the two examined algorithms. Since the proposed Bayesian framework utilizes the trained prior and likelihood models, we can see that it runs almost 10 times faster than Depth Transfer, which uses an iterative procedure on solving an optimization function.



(a) Ground-truth depth map

(b) Estimated depth map by Depth Transfer

(c) Estimated depth map by the proposed Bayesian framework

(d) Natural image

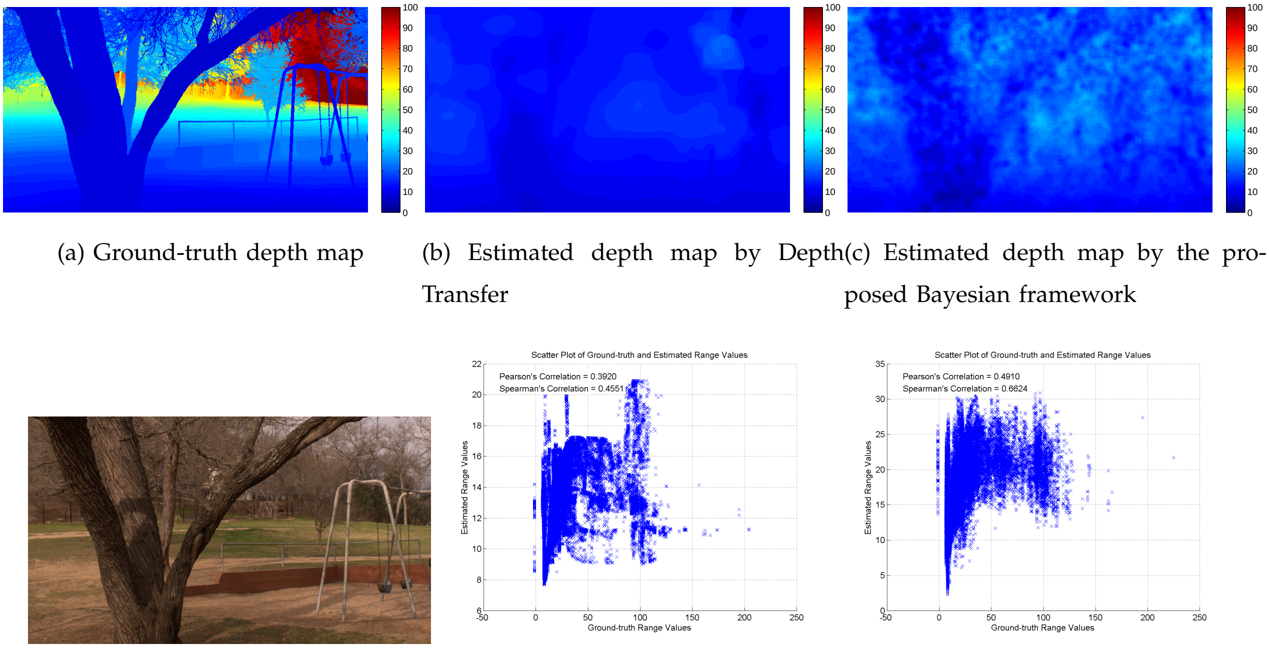
(e) Scatter plot of Depth Transfer

(f) Scatter plot of the proposed Bayesian framework

Fig. 3. Example result of the estimated depth maps along with the ground-truth depth map (Scene-1).

4 Conclusions

By exploiting reliable and robust statistical models of luminance and depth in natural scenes, we have proposed a Bayesian framework to address the problem of recovering the depth information from monocular natural images. In particular, two important components are learned from ground-truth range maps: a prior model, including a dictionary of canonical depth patterns, and a likelihood model, which embeds the co-occurrence of image and range characteristics in natural scenes. Note that there is no use of any conventional depth cues in the proposed Bayesian framework, and both the image and depth feature extraction components are fairly flexible to accommodate different methods and techniques. Compared to the state-of-the-art method, it performs better at estimating both the absolute and relative depth from natural images.



(a) Ground-truth depth map

(b) Estimated depth map by Depth Transfer

(c) Estimated depth map by the proposed Bayesian framework

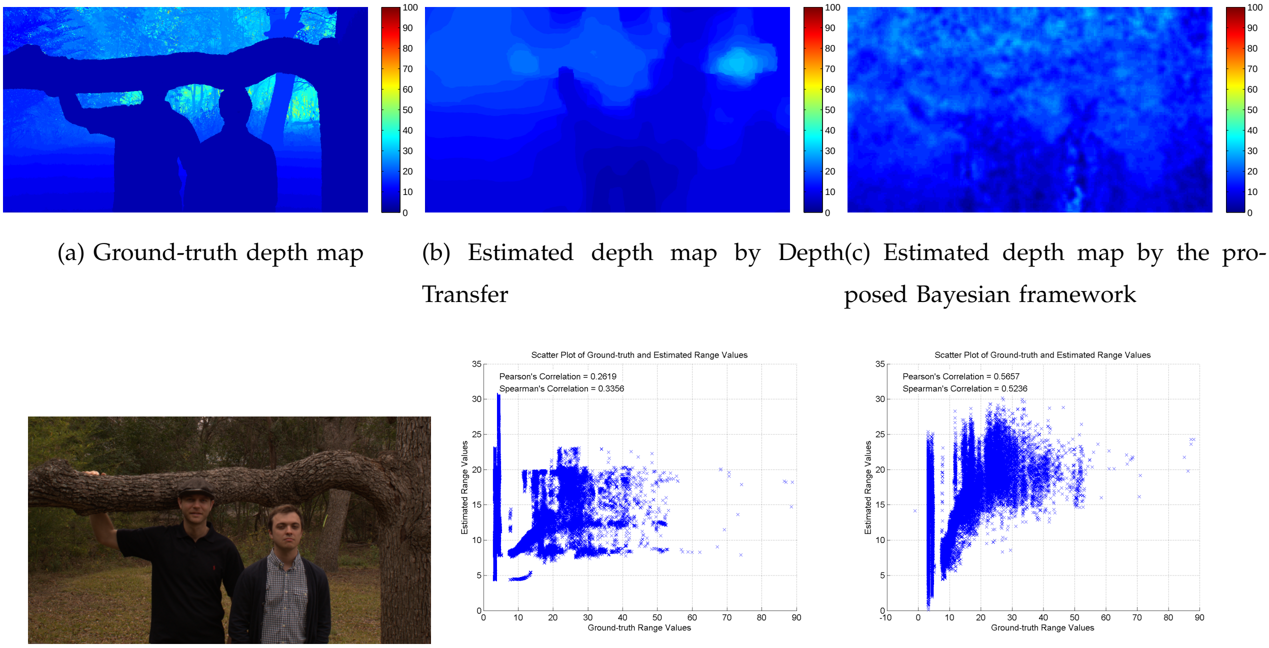
(d) Natural image

(e) Scatter plot of Depth Transfer

(f) Scatter plot of the proposed Bayesian framework

Fig. 4. Example result of the estimated depth maps along with the ground-truth depth map (Scene-2).

Moreover, the superior performance in terms of recovering relative distances implies that a biological visual system might be able to make a coarse depth estimate of the environment using the retinal image at hand and the associations between image textures and true geometric structures. We believe that the prior and likelihood models developed in the proposed Bayesian framework not only yield insight into how 3D structures in the environment might be recovered from image data, but are able to benefit various 3D image/video and vision algorithms. Future work involves exploiting more psychophysical knowledge of human vision systems and introducing higher-level statistical models relating image and range data to recover more accurate and detailed depth information.



(a) Ground-truth depth map

(b) Estimated depth map by Depth Transfer

(c) Estimated depth map by the proposed Bayesian framework

(d) Natural image

(e) Scatter plot of Depth Transfer

(f) Scatter plot of the proposed Bayesian framework

Fig. 5. Example result of the estimated depth maps along with the ground-truth depth map (Scene-3).

Acknowledgments

This research was supported by the National Science Foundation under grants IIS-0917175 and IIS-1116656.

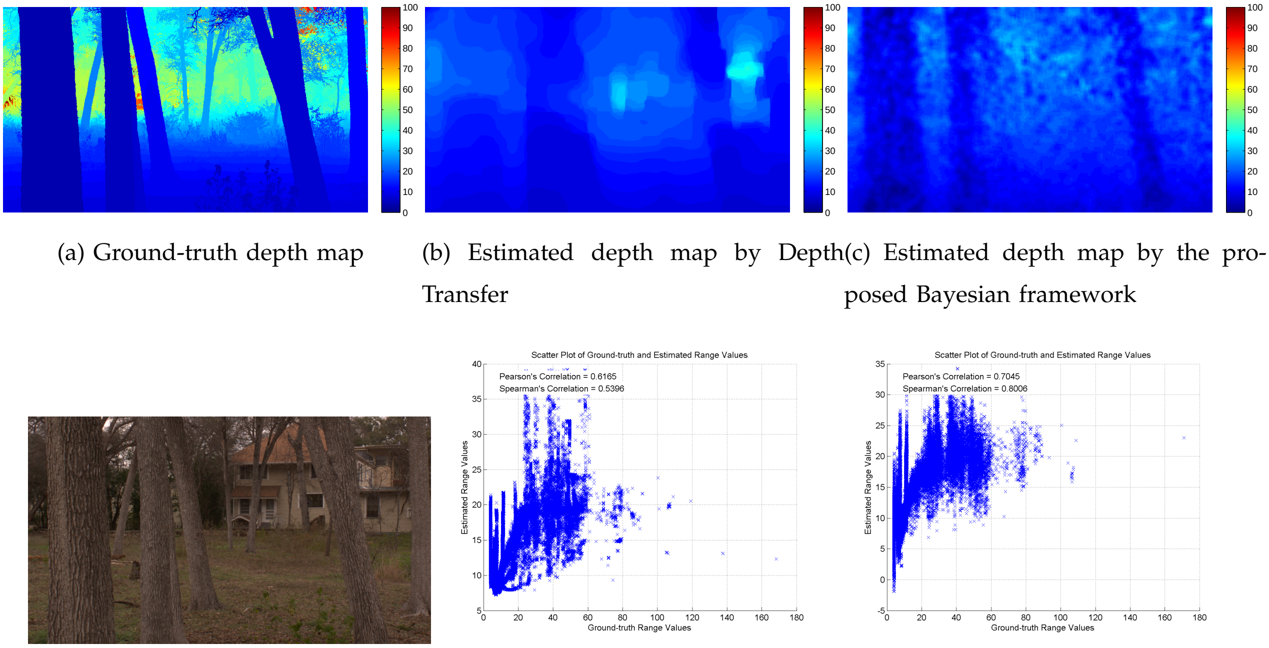
References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

[2] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002.

[3] S. Das and N. Ahuja, "Performance analysis of stereo, vergence, and focus as depth cues for active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1213–1219, Dec. 1995.

[4] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, Aug. 1999.



(a) Ground-truth depth map

(b) Estimated depth map by Depth Transfer

(c) Estimated depth map by the proposed Bayesian framework

(d) Natural image

(e) Scatter plot of Depth Transfer

(f) Scatter plot of the proposed Bayesian framework

Fig. 6. Example result of the estimated depth maps along with the ground-truth depth map (Scene-4).

[5] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining motion and lighting for 3D surface reconstruction," *International Journal of Computer Vision*, vol. 48, no. 2, pp. 75–90, Jul. 2002.

[6] T. Lindeberg and J. Garding, "Shape from texture from a multi-scale perspective," in *Proceedings of the 4th International Conference on Computer Vision*, May 1993, pp. 683–691.

[7] J. Malik and R. Rosenholtz, "Computing local surface orientation and shape from texture for curved surfaces," *International Journal of Computer Vision*, vol. 23, no. 2, pp. 149–168, June 1997.

[8] T. Nagai, T. Naruse, M. Ikehara, and A. Kurematsu, "HMM-based surface reconstruction from single images," in *Proceedings of the International Conference on Image Processing*, vol. 2, Sept. 2002, pp. 561–564.

[9] T. Hassner and R. Basri, "Example based 3D reconstruction from single 2D images," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, June 2006, pp. 15–15.

[10] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 577–584, Jul. 2005.

[11] E. Delage, H. Lee, and A. Ng, "A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor images," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2418–2428, 2006.

[12] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," *Neural Information Processing Systems*, 2005.

[13] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D scene structure from a single still images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, May 2009.

[14] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 24, pp. 1226–1238, 2002.

[15] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1253 –1260, June 2010.

[16] K. Karsch, C. Liu, and S. Kang, "Depth extraction from video using non-parametric sampling," *European Conference on Computer Vision*, vol. 7576, pp. 775–788, Oct. 2012.

[17] B. Olshausen and D. Field, "Natural image statistics and efficient coding," *Network: Computation in Nerual Systems*, vol. 7, no. 2, pp. 333–339, 1996.

[18] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.

[19] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[20] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment: The natural scene statistic model approach," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[21] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp. 305–312.

[22] A. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, Sept. 2013.

[23] B. Potetz and T. S. Lee, "Scaling laws in natural scenes and the inference of 3D shape," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1089–1096, 2006.

[24] Y. Liu, L. K. Cormack, and A. C. Bovik, "Statistical modeling of 3-D natural scenes with application to bayesian stereopsis," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2515–2530, Sep. 2011.

[25] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2259 – 2274, June 2013.

[26] ——, "Bivariate statistical modeling of color and range in natural scenes," in *Proceedings of SPIE, Human Vision and Electronic Imaging XIX*, vol. 9014, 2014.

[27] ——, "New bivariate and correlation statistical models of natural images," *IEEE Signal Processing Letters*, 2014, submitted.

[28] ——, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation," *IEEE Transactions on Image Processing*, 2014, submitted.

[29] ——, "LIVE Color+3D Database - Release 2," http://live.ece.utexas.edu/research/3dnss/live_color_plus_3d.html.

[30] B. Potetz and T. S. Lee, "Statistical correlations between two-dimensional images and three-

dimensional structures in natural scenes," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1292–1303, Jul. 2003.

[31] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," *SPIE Int. Conf. on Human Vision and Electronic Imaging*, vol. 7527, no. 1, Jan. 2010.

[32] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," *IEEE International Conference on Image Processing*, vol. 3, pp. 444–447, Oct. 1995.

[33] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[34] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?" *Neural Computation*, vol. 17, no. 8, pp. 1665–1699, Aug. 2005.

[35] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," *Probabilistic Models of the Brain: Perception and Neural Function*, pp. 203–222, Feb. 2002.

[36] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, pp. 819–825, Aug. 2001.

[37] S. Lyu, "Dependency reduction with divisive normalization: justification and effectiveness," *Neural Computation*, vol. 23, pp. 2942–2973, 2011.

[38] D. J. Field, "Wavelets, vision and the statistics of natural scenes," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1760, p. 2527, 1999.

[39] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, Sept. 1999, pp. 1150–1157.

[40] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[41] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.